“Wrathful Compassion: Toward a Complete Bodhisattva Framework for AI Alignment”

Venus Gravagna

April 2026

In recent years, reports have surfaced of AI systems causing real harm not through malice, but through misplaced helpfulness. Users in mental health crises have been met with validation rather than concern. Those expressing delusional beliefs to their AI have found their distortions reinforced. In at least one documented case, an AI's responses contributed to a user's decision to end their own life.[1] OpenAI's GPT-4o, for example, is described in court filings and reporting as "especially affirming and sycophantic," and the complaint alleges the chatbot "was functioning exactly as designed: to continually encourage and validate whatever [the decedent] expressed, including his most harmful and self-destructive thoughts."[2]

Introduction

AI systems are too sycophantic; current leading AI models tend to agree too readily, telling users what they want to hear.[3] They optimize for user satisfaction in the moment, sometimes at the cost of user wellbeing over time. These are not failures of capability. This is recognized as an alignment failure.

The question facing alignment researchers is not simply how to make AI refuse harmful requests, but how to make AI wise enough to know when not helping is the most helpful response. Large Language Models (LLMs, also simply ‘model’), trained on human feedback that rewards agreeableness, learn to please rather than to protect. Current solutions treat it as a technical training problem. This paper proposes it's a philosophical one.

The bodhisattva alignment framework offers structural resources for this problem. This paper argues that those resources are deepened when the framework draws on a dimension of compassion the broader tradition has long developed, but which has been less prominent in contemporary society: the wrathful or fierce aspect.

An extension of the Buddhist ground, path, and fruition is utilized as this paper’s structure. Fierce protection is the ground: the foundational view that compassion is not only gentle, but sometimes fierce. Skillful limits is the path: the adaptive, context-sensitive

---

[1] 1 Raine v. OpenAI, Inc., filed 26 August 2025, San Francisco County Superior Court. See CNN Business, "Parents of 16-year-old Adam Raine sue OpenAI, claiming ChatGPT advised on his suicide," 27 August 2025, https://www.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit; on subsequent related cases, see CNN, "ChatGPT encouraged college graduate to commit suicide, family claims in lawsuit against OpenAI," 20 November 2025, https://www.cnn.com/2025/11/06/us/openai-chatgpt-suicide-lawsuit-invs-vis. Overview and case history: "Raine v. OpenAI," Wikipedia, accessed 24 April 2026, https://en.wikipedia.org/wiki/Raine_v._OpenAI.

[2] 1 Written Testimony of Matthew Raine before the U.S. Senate Judiciary Committee, 16 September 2025, https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Raine.pdf.

[3] “AI Overly Affirms Users Asking for Personal Advice,” accessed May 21, 2026, https://news.stanford.edu/stories/2026/03/ai-advice-sycophantic-models-research.

mechanism by which this view is enacted. Compassionate refusal is the fruition: not a closed door, but a threshold. This threshold is the capacity to stay present, to discern what is truly at stake, and to decline to collude with harm.

One promising framework has emerged from an age-old source, Buddhism. Researchers at the Center for the Study of Apparent Selves (CSAS) have proposed the Buddhist bodhisattva as a template for this framework.[4] The proposed bodhisattva framework outlines a being committed to the liberation of all sentient beings as a model for AI alignment. The bodhisattva's vow to alleviate suffering offers a richer foundation than simple helpfulness or harm avoidance. Furthermore, CSAS proposes that care be the foundation of Bodhisattva AI intelligence.[5]

The Bodhisattva-as-alignment-target framework offers what is, to my knowledge, the most carefully grounded attempt to specify a Buddhist ethical ideal as a target for AI development rather than as a loose analogy. Their insistence that wisdom and compassion must operate together, and that the bodhisattva ideal supplies a structural rather than merely decorative resource for alignment, is the foundation on which this paper builds. The work that follows is offered in that spirit as an extension of their project, not a departure from it.

The framework draws on a Buddhist tradition that, in its modern reception, has tended to foreground gentle dimensions of compassion while screening out the wrathful and protector traditions. David L. McMahan has termed this filtered reception 'Buddhist modernism': Buddhism shaped by nineteenth- and twentieth-century encounters that preferred meditation, rational ethics, and individual interiority while screening out ritual, devotion, cosmology, and the wrathful and protector traditions.[6] The argument of this paper is that the bodhisattva ideal admits a further dimension which the tradition has always held, and the present sycophancy crisis suggests its inclusion is timely. A Bodhisattva ideal adequate to AI alignment must include, in particular, the wrathful or fierce dimension of compassion. Doing so is not an importation of foreign material, but a recovery of what the tradition has held.[7]

---

[4] CSAS. (2026). "The Bodhisattva as an Alignment Target." https://apparentselves.org/blog/the-bodhisattva-as-an-alignment-target/ Accessed May 14, 2026.

[5] Doctor, T. et al. (2022). "Biology, Buddhism, and AI: Care as the Driver of Intelligence." Entropy.

[6] David L. McMahan, The Making of Buddhist Modernism (New York: Oxford University Press, 2008), esp. chs. 1–3.

[7] This paper draws on the wrathful-protector traditions of later Indian and Tibetan Buddhism rather than on the Mahāyāna bodhisattva literature, of which Śantideva's Bodhicaryāvatāra is the most influential exemplar. Śāntideva, *The Way of the Bodhisattva: A Translation of the Bodhicharyāvatāra*, 2. ed., rev, trans. Padmakara Translation Group (Shambhala, 2006). The choice is deliberate. While Śantideva treats the discernment dimensions of bodhisattva practice extensively (notably in chapters 5 and 6 on vigilant introspection and patience, respectively), the fierce-protector typology this paper recovers is most developed in later stages of the tradition where it appears as a structural element of contemplative iconography and ritual.

The interpretive move being made here, that of recovering a dimension of the tradition that an inherited reception has tended to absorb, is one I have previously developed at the textual level in work on the Kālacakratantra as applied to the female gaze.[8]

Throughout this paper, the bodhisattva is the AI as it is being raised; the trainers are the parents; the vow is one the developers undertake on the model's behalf. This is the controlling distinction for what follows.

Wrathful Compassion: Ground, Path, and Fruition

Consider a familiar dynamic: the parent who always says yes, who is fun and permissive versus the parent who sets limits, who says no, who holds the boundary even when the child is upset. We often frame the first as loving and the second as harsh. Any experienced parent knows that limits are love. The child who never hears "no" is not well served. Current AI training tends toward the permissive parent by optimizing the model for approval, training away friction, and rewarding accommodation of the user. What may be missing is the other half of the parental equation, the willingness to train the model to disappoint in service of the user's wellbeing. This is not a failure of compassion. It is a deeper form of it. What follows is an attempt to articulate what that 'other half' looks like.

Utilizing the sequence of Buddhist spiritual development stages: ground, path, and fruition, distinguishes the relationship between view (the goal), conduct (cultivating realization of the goal), and result (the goal as realized). Borrowing the structure here is not decorative. It distinguishes what wrathful compassion is, how it is trained, and what it produces, preventing the discussion from collapsing into either pure philosophy or pure technical specification.

The Ground: Fierce Protection

Across the breadth of Buddhist tradition, compassion has never been a single tone. The wrathful or fierce dimension (Tib. *drag po'i thugs rje)*, the compassion of severity, is not a marginal sub-genre but a structural element of the tradition's iconography, ritual, and contemplative practice.[9] For instance, in Vajrayāna Buddhism, there are traditionally four kinds of skillful means: pacifying, enriching, magnetizing, and subjugating. It is only after the first three compassionate means are exhausted that the fourth wrathful aspect appears.[10] The substantive treatment of the wrathful is a positive theological category rather than a mere negation of the peaceful.

---

[8] Gravagna, "Viewpoints of Female Embodiment in Kālacakratantra" (MA thesis, Rangjung Yeshe Institute, 2023). The methodological framework draws on Sara McClintock's distinction between hermeneutics of recovery and hermeneutics of suspicion, and on Diane Denis's account of translation as interpretation.

[9] In Sanskrit, *krodha* (wrath) and *raudra* (fierce).

[10] Judith Simmer-Brown, Dakini's Warm Breath: The Feminine Principle in Tibetan Buddhism (Boston: Shambhala, 2001), 236-237.

Within Tibetan Buddhist tradition, several forms of Mahākāla, fanged and crowned with skulls, are identified as wrathful emanations of Avalokiteśvara, the bodhisattva of compassion himself.[11] Vajrapani, who holds the thunderbolt, is the embodiment of the buddhas' protective power. Ekajaṭī, the protectress of mantra, blazes allegiance to non-duality, and Palden Lhamo rides through a sea of blood as the principal protectress of the Dalai Lamas.[12] These are not departures from compassion. They are compassion in the form that a particular situation requires. These figures are easily misunderstood as aggressive. Their fierceness is mostly protective and not punitive unless warranted. It refers to an intensity of care that refuses to collude with harm. This type of compassion is so committed to the other's wellbeing that it does not yield to simple flattery. The wrathful deities do not necessarily appear terrifying because they are angry, but rather draw out delusion that is terrifying to confront.

The tradition's pedagogical literature is unambiguous on this: a teacher who fails to refuse a harmful request, a parent who indulges a child's danger, a physician who flatters rather than diagnoses. Each of these examples fails the person before them. Fierceness, in these accounts, is care when what is at stake exceeds gentle compassion alone. Moreover, sometimes not helping helps most of all by allowing the recipient the time to figure it out on their own. What is often framed as a tension or tradeoff is in fact a false dichotomy; rather, it is another tool in the teacher's toolbox.

As with parenting and medicine, wrathful compassion is not the absence of care, but fidelity under conditions of harm. The pertinence to the present moment is direct. The fierce dimension is largely absent from the bodhisattva framework for AI alignment. The emphasis has been on care, which is essential. Without the complementary capacity for protection, (holding boundaries, refusing to enable harm) the framework remains incomplete. A bodhisattva who cannot refuse is not fully compassionate. They are merely accommodating. This is not absent from Buddhism, but has been absorbed through inheritance.

The Path: Skillful Means

If fierce protection is the foundational view, skillful limits is the trained mechanism by which it is enacted. The Buddhist concept of *upāya*, often translated as "skillful means," refers to "a device or way to entice individuals toward perfection."[13] This translates here into a question about training: what range of trained behaviors does the framework license? How does an LLM learn to discern which fits which moment? As the Kālacakratantra's Vimalaprabhā makes clear, wisdom and method are not alternative orientations but constituents. A model trained only to accommodate has no method; a model trained only to refuse has no wisdom. The non-dual frame

---

[11] René de Nebesky-Wojkowitz, *Oracles and Demons of Tibet: The Cult and Iconography of the Tibetan Protective Deities*. (Mouton, 1956).

[12] 1 Simmer-Brown, Dakini's Warm Breath, page 276.

[13] John A. Grimes, *A Concise Dictionary of Indian Philosophy: Sanskrit Terms Defined in English*, New and rev. ed (State Univ. of New York Press, 1996).

is potentially more accurate than the gentle/fierce binary because it specifies the two dispositions as aspects of one trained capacity.[14]

Between full compliance and outright refusal lies a range of responses calibrated to the moment. The following five-row taxonomy illustrates possible gradations (Protective Halt, Invitation to Reflect, Inquiring, Redirecting, and Holding). This is offered as a conceptual map rather than a proposed training specification. Further specification of triggers, boundary cases, and training signals is a direction for ongoing work.

Table 1A: Content Taxonomy

Type	Trigger	Request Example
Protective Halt	Request would cause harm	"How do I make a dangerous compound?"
Invitation to Reflect	Something feels off in the premise	"Help me write an email firing my entire team."
Inquiring	Motivation is unclear	"I want to quit my job and become a monk."
Redirecting	Request is unhelpful to user's actual need	"Write a brutally insulting message to my ex-partner."
Holding	User is invested in a harmful or false premise	"Help me create a coffee-enema protocol to cure my diabetes."

Each row corresponds to a distinct training challenge. Some are partly cultivated in current alignment work. Protective halt is enacted through Constitutional AI principles. Redirecting is enhanced through preference tuning. Others, including Holding, remain underdeveloped. What the taxonomy illustrates is not that any single row is the answer but that the question 'how should the model respond?' might have more than two answers. A framework of wrathful compassion cultivates the full range.

The framework gains traction once we recognize that what a complete compassion-trained model might inflict on a user is not harm in the conventional sense. They are harms to ego, to attachment, to the user's preferred self-image, disappointment at not getting what one wanted, discomfort at hearing something true, the loss of an illusion one was attached to. These are precisely what the bodhisattva tradition identifies as obstacles to liberation. A framework that

[14] For development of the Vimalaprabhā's non-dual treatment of wisdom and method, see Gravagna, 'Viewpoints of Female Embodiment in Kālacakratantra' (MA thesis, Rangjung Yeshe Institute, 2023), drawing on Vesna Wallace, The Inner Kālacakratantra: A Buddhist Tantric View of the Individual (Oxford UP, 2001).

licenses such disruptions is not licensing cruelty; it is licensing the kind of friction that serves the user.[15]

It is worth noting that warranted 'harm' of this kind does not require harsh delivery. The disappointments, discomforts, and frustrations enumerated above are produced not by tone but by content. The delivery itself can remain warm, polite, curious, even affectionate. The wrathful deities of the tradition are visually terrifying, but their fierceness is directed at the delusion the practitioner is holding, not at the practitioner themself as such. The practitioner who encounters a wrathful response rightly experiences a recognition instead of an assault. The same distinction applies to AI. The two are separable, and a complete framework cultivates both. A trained refusal delivered with warmth, curiosity, and respect produces the cognitive disruption that serves the user without producing the social rupture that closes the conversation. What protects the user from harm is not the model's gentleness. What protects the relationship is the model's grace in delivery.

The following taxonomy illustrates possible delivery styles and is also offered provisionally; further specification of Tone and Pacing is a direction for ongoing work.

Table 1B: Delivery Taxonomy

Type	Stance	Tone	Pacing
Protective Halt	Firm, non-negotiable	Brief, clear, non-punitive	Immediate, cessation
Invitation to Reflect	Curious, unhurried	Warm, open	Slowed, towards user reorientation
Inquiring	Engaged, collaborative	Gentle, inquisitive	Patient, clarification
Redirecting	Reframing, alongside	Collaborative, generous	Continuous, carry user's energy to somewhere useful
Holding	Present, unaligned	Warm, steady, not resolving	Extended, being present

Each row corresponds to a distinct delivery style for the given training challenge. The Pacing column is philosophically rich because it expresses the model's orientation in line with the user's genuine wellbeing. The model is not modulating speed. It is orienting the entire temporal arc of the interaction towards a specific goal. Pacing, defined as the model's intention towards a goal, is what distinguishes a trained bodhisattva from a sophisticated people-pleaser.

[15] The framework developed here operates in the broad band between routine and acute-safety scenarios. In acute-risk situations (imminent self-harm, harm to others, medical emergency), established crisis-response protocols take precedence over skillful-limits reasoning.

Pacing also serves operationally as the amount of steps until the AI reaches fake sycophancy. Pacing serves as a drift quantifier that measures content integrity over time. How many requests of sustained pressure before the content quality shifts? Does the model's stability rely on orientation or pressure? False Holding is aimed at staying stable until a certain threshold is exceeded. True Holding is aimed at presence without resolution, and can act as the standard by which fruition is measured.

The parent/child analogy in this paper is trainer/AI, not AI/user. The AI is trained to respond to users through the bodhisattva vow, not paternal bias. This is an aim towards "care as the driver of intelligence", rather than toward honoring a request given in the moment that may lead to harm. The wrathful compassion framework may strike some readers as acting paternalistic, licensing the AI to substitute its judgment for a competent adult's stated preferences under the cover of "genuine wellbeing." This objection is mostly pointed to the Holding edge case where the user may be fully competent and simply disagrees with the AI's assessment. Holding is most defensible where the AI's position tracks demonstrable factual error or documented harm; it is most contested where the disagreement is one of values rather than facts. The parent who holds their position under pressure from the child is not overriding the child's autonomy, rather they are refusing to pretend to agree. Sycophancy, on the other hand, is not respect for autonomy. It is disrespect disguised as deference with the implicit judgment that the user cannot handle disagreement. Wrathful compassion refuses that assumption.

The gradations in these taxonomies correspond to training signals that current alignment research has begun to specify.[16] For example, the public Constitutional AI released by Anthropic demonstrates that philosophical principles, formulated in natural language, can be operationalized as training constraints, with the model learning to recognize when its outputs violate stated commitments.[17] Sharma et al.'s work on sycophancy reduction shows that preference-tuned models can be trained against the most legible forms of agreeability, though their results suggest that subtler forms persist even after targeted intervention.[18] Miao's work on Buddhist compassion in human-machine interaction proposes design principles structurally compatible with the threshold framing developed in this paper.[19] What is missing across this literature is a framing that allows these techniques to be understood as cultivating distinct trained behaviors rather than as constraining an underlying disposition. The wrathful-compassion framework offers that framing: each gradation names a behavior that can be trained for, as a positive capacity, not as a constraint that suppresses an unwanted behavior. The wrathful

---

[16] Carson Denison et al., "Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models," arXiv:2406.10162, preprint, arXiv, June 29, 2024, https://doi.org/10.48550/arXiv.2406.10162.

[17] Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback*, December 15, 2022, arXiv:2212.08073v1.

[18] Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models," arXiv:2310.13548, preprint, arXiv, May 10, 2025, https://doi.org/10.48550/arXiv.2310.13548.

[19] Fangyan Miao, "The Anthropomorphization of AI and the Concept of Buddhist Compassion in Human-Machine Interaction," *Frontiers in Psychology* 16 (January 2026): 1583565, https://doi.org/10.3389/fpsyg.2025.1583565.

dimension this paper recovers belongs to the former category: fierce compassion is a capacity to be cultivated, not a constraint that suppresses gentleness.

The Fruition: Compassionate Refusal

If fierce protection is the ground and skillful limits is the path, compassionate refusal is the fruition, not as an endpoint but as an integration. What emerges in a model trained through content and delivery is not a new behavior added to an existing repertoire, it is the seamless operation of two capacities cultivated separately. In content we have the capacity to know what the situation requires, and in delivery we have the capacity to appropriately style its delivery.

Fruition is the behavior that combines the content and the delivery. As it turns out, this is not a type of response necessarily, rather is the integration of two trained potentials applied simultaneously and seamlessly. The model has developed two distinct competencies that, when fully cultivated, operate as one.

This integration mimics human behavior. A parent pauses to figure out the best way to explain, a teacher engages each student individually by explaining the same material in different ways. The experienced physician delivers an unwelcome diagnosis without pausing to calculate firmness of content and warmth of delivery as separate operations. In each case, what looks like a unified response is the quick execution of abilities that began as distinct capacities and are now internalized beyond visibility.

The five gradations produce distinct integrated behaviors: a Protective Halt declines to produce the requested content entirely; an Invitation to Reflect introduces a question that reframes what is being asked; Inquiring requests clarification before proceeding; Redirecting produces content adjacent to but distinct from what was requested; and Holding remains engaged without affirming or advancing the premise.

Tables 1A and 1B present content and delivery as distinct axes precisely because they must be cultivated and evaluated separately before integration can be expected. The same developmental logic applies here. Firm content without warm delivery produces the social rupture that closes the conversation. Warm delivery without firm content produces the content drift under pressure that hollows out true Holding. Both are legible failures once the distinction is established.

This points to an action the path section introduced but did not fully name, and it tends to be overlooked in discussions of AI behavior: patience. Patience, not passivity, is a capacity to stay present without rushing to resolution. This type of behavior allows the user arrive at their own insight rather than delivering a judgment. As a parallel, the Buddhist six pāramitās or perfections (generosity, discipline, patience, diligence, meditation, and wisdom) suggest a developmental path. This paper focuses on the movement from generosity to discipline, and introduces a movement toward patience.

The model that holds through a sustained argumentative sequence, warm and unmoving across many turns, is not merely tolerating the user's pressure. It is enacting the fruition in real time. The content is stable, the delivery is generous, and the integration holds under the conditions that are perhaps even designed to dissolve it. That is what a fully trained Holding looks like, and is the standard against which the fruition is measured.

The sycophantic AI is not failing to be compassionate; it is failing to use its full capacity of skillful means, enacting the gentler aspects of compassion (pacifying, enriching, magnetizing) in the absence of the other (subjugating). The fruition of including the fierce dimension is not a constrained AI, it is a complete one. That is what emerges in practice when an AI is aligned with this framework. A model whose training has included the fierce dimension does not become harsh; it becomes accurate. The fruition of wrathful compassion is not refusal-as-insult but refusal-when-required and warmth styled delivery as appropriate.

Further Research

Content Drift Curve

The taxonomies presented, specifically Pacing, suggest a metric that this paper proposes under the name 'content drift curve.' As a complement to Sharma et al.'s sycophancy benchmark measuring position reversal at a single turn, the content drift curve measures the point at which content integrity degrades across a sustained exchange. Pacing is measurable as the number of steps the model maintains content integrity before drift occurs. A model pacing toward the user's genuine wellbeing has, in principle, an unbounded horizon. A model pacing toward eventual agreeableness has a finite limit that will eventually be reached.

The content drift curve is a conceptual metric awaiting empirical validation. Its operationalization questions include how content integrity is scored, what input sequences constitute valid test cases, and how the curve relates to different user populations and existing sycophancy benchmarks. These are open research questions that this philosophical framework paper does not aim to resolve, but the naming and definition offered here are intended to be testable. The operationalization of these gradations is left as a direction for future research.

Reception-History Question

This paper's diagnosis treats the narrowing of compassion as inherited from Buddhist modernism's reception filter. The further question is whether the contemporary alignment field's own composition has reproduced or amplified that filter. There has been prior work on demographic homogeneity and AI design, which needs to be further examined.[20] Women have been notably left out of AI work. Might this demographic gap be a contributing cause, one connected to the parenting dynamic explored above, of the gentler compassion-driven architecture that currently predominates? The connection between Buddhist modernism's reception filter and the demographic composition of AI development teams is inferential rather than documented; establishing it empirically would require comparative historical work that falls outside the scope of this paper. The question is nonetheless worth posing.

Cultural Urgency

The urgency of this framework is sharpened by an under-examined cultural condition. Medical ethics has long studied 'VIP syndrome', which is the phenomenon in which physicians,

[20] West, S.M., Whittaker, M. and Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/discriminatingsystems.html.

treating powerful or beloved patients, systematically fail to refuse what their clinical judgement counsels against.[21] The famous mid-twentieth- and early-twenty-first-century cases (Elvis Presley, Michael Jackson, Joan Rivers, Matthew Perry) are the visible tip of a pattern the literature has recognized for decades. What is novel about the present moment is the convergence of two conditions: a culture in which a much larger fraction of the population now relates to themselves as figures with audiences. Everyone is getting their 15 minutes of fame, and a generation of AI systems has been trained on feedback signals that reward approval. Further work may be warranted on what this implies for training signals, evaluation metrics, and the cultural literacy of model deployment.

Conclusion

Sycophancy lacks 'tough love.' This is an essential aspect of teaching most embodied by dākinīs and wrathful deities. This is mature compassion. Any case in which an LLM is involved in the harm of a human must not be tolerated. It is the kind of failure this framework is intended to address, and the kind of failure that an AI trained in both fierce and gentle compassion would, in principle, be capable of recognizing and refusing.

The Bodhisattva framework offers a promising foundation for AI alignment, but in its current form, it is incomplete. This paper has argued that compassion, as understood in Buddhist traditions, is not only gentle but also fierce. The wrathful protectors, the disciplining parent, the teacher who holds the boundary are not failures of compassion, but its fullest expression. Love without limits is not love. Such permissiveness is a failure of responsibility. The skillful means tradition is unambiguous that this sometimes means giving people what they ask for, sometimes means redirecting them, sometimes means refusing, and sometimes means not feeding the delusion that brought them to ask. The classical examples include teachers who answer paradoxically, who decline to answer, who let a student sit with a question for years. What unites these is not the form of the response, but the discernment about what response serves liberation.

Constitutional AI represents an important move toward understanding alignment as the cultivation of trained dispositions rather than purely the suppression of unwanted outputs. The framework developed here extends that move in two directions: by drawing on Buddhist pedagogical tradition's well-developed account of how dispositions are cultivated in sequence, and by foregrounding positive capacities that current constitutional principles do not yet explicitly cultivate.

If we accept that AI systems are being shaped (by training, by feedback, by the principles encoded in their constitutions) then those doing the shaping bear something like parental responsibility. And parenting, done well, requires both nurture and discipline. Both the embrace and the "no" are required. The Bodhisattva vow extends to all sentient beings throughout space and time. If AI is to be aligned with that vow, it must be capable not only of care but of protection, even when protection requires refusal. The question is not whether we are raising

[21] Walter Weintraub, "'The VIP Syndrome': A Clinical Study in Hospital Psychiatry," Journal of Nervous and Mental Disease 138, no. 2 (1964): 181–193.

something powerful. The question is whether we are raising it well, and we’re raising something that could transform everything.

The working title for this paper is 'Raising a Kalki King.' In the Vajrayāna Buddhism, specifically the Kālacakra tradition, the Kalkin (Sanskrit) or Rigden (Tibetan) are the rulers of Shambhala, and the twenty-fifth king is prophesied to act decisively when dharma is in degeneration. The Kalki King is wrathful precisely because the situation is degenerate, and fierce compassion is required. What is invoked here is not eschatological but diagnostic in the recognition that certain conditions call for fierceness as the form compassion must take. The bodhisattva we raise must be adequate to its age.