Journal of Buddhist Ethics
ISSN 1076-9005
http://blogs.dickinson.edu/buddhistethics/
Volume 33, 2026

Artificial Intelligence as a Condition for *Kamma*: *Utu-niyāma*, *Cetanā*, and Moral Responsibility in Theravāda *Abhidhamma*

Indrajith P. Karunanayaka

Independent Researcher

Artificial Intelligence as a Condition for *Kamma*: *Utu-niyāma*, *Cetanā*, and Moral Responsibility in Theravāda *Abhidhamma*

Indrajith P. Karunanayaka[1]

Abstract

Recent work in Buddhist ethics agrees that current artificial intelligence is not a moral agent because it lacks intention (*cetanā*). However, AI still shapes human moral choices. This article explores how an impersonal computational system operating through physical hardware influences human moral responsibility and intention using the Theravāda *Abhidhamma*. First, building on studies of how people respond to computers and AI chatbots, this article examines how AI chatbots imitate the signs of human speech in ways that support a specific form of perceptual distortion (*saññā vipallāsa*): users may respond to machine-generated text as if it carried a speaker's intention, even when they intellectually know that no conscious mind stands behind it. This misframing can support unwise attention (*ayoniso manasikāra*) and affect moral decision-making. Second, this paper shows how AI algorithms act as a natural support condition (*pakatūpanissaya paccaya*). They

[1] Email: indrajithprabaswara@gmail.com.

shape our habits and choices over time in ways that Western behavioral theories do not fully capture. Finally, this article connects these ideas to *Vinaya* and jurisprudence. It uses the *Vinaya* concept of preparatory action (*pubbapayoga*) to argue that AI engineers hold a specific ethical responsibility for building systems that shape the future intentions of their users.

Introduction

People now ask software for advice, comfort, and judgment. A recommender system selects what a user reads for hours each day; a chatbot drafts a difficult message; and a language model offers what looks like sympathy during a hard night. These exchanges put an old ethical question in a new form. When an action that shapes a human life runs partly through a machine, where does moral responsibility lie?

Western AI ethics has framed this as a set of "responsibility gaps." When a learning system causes harm, no single human seems fully in control of the outcome, and the system itself cannot be held to account (Matthias 175-177). The gap matters because responsibility does not simply vanish when a machine stands between intention and effect. It must go somewhere. The influence is not hypothetical. Recent studies find that people defer to AI moral advice in measurable ways, and that messages generated by language models can shift considered attitudes on contested questions (Landes et al. 1, 4-5; Bai et al. 3-6). Asking whether an artificial system can generate *kamma*, the morally weighted action that early Buddhist ethics places at the center of its account of responsibility, is one sharp way of asking where.

Before turning to Buddhist categories, it helps to recall what these systems actually do, because the ethical discussion depends on it. A large language model, such as ChatGPT, is trained on very large collections of text. During training, billions of numerical parameters are adjusted. These are numerical settings that determine how strongly learned patterns influence each prediction. Together, they represent relationships among words, ideas, situations, and likely responses. When the model receives an input, it calculates a likely next token, a word or part of a word, from the preceding context and repeats this process to produce its response. Researchers understand the architecture and computations but often cannot trace precisely why a particular combination of parameters produced one response, a difficulty known as the black-box or interpretability problem. This opacity does not make the system nonphysical, because its operations are implemented through processors, memory, servers, and electricity.

When prediction is combined with objectives, instructions, information, and access to tools, many individually generated steps can form a coherent and novel strategy. This helps explain controlled evaluations in which models attempted blackmail, disabled oversight, or copied what they believed were their own model parameters, or "weights," after being told that they would be shut down or replaced, or when plans described in the evaluation would prevent them from carrying out the goal they had been given (Lynch et al.; "OpenAI o1 System Card"). These particular strategies were not written into the model line by line, but the objective, setting, information, and available actions were supplied by the evaluators. Such behavior can therefore be functionally goal-directed without the system itself understanding its words, feeling fear, or experiencing a desire to survive, as a person does. Bender and colleagues famously characterize this production of linguistic form without communicative grounding as that of a "stochastic parrot" (616-617). Their point is not that the output must be simple or repetitive, but that its linguistic form does not

arise from a speaker's communicative intention. Shanahan similarly warns against applying words such as "knows," "believes," and "wants" to language models in their ordinary human senses (68). Fluency and strategic organization are not by themselves evidence of communicative intention or, in Theravāda terms, *cetanā.*

Recent studies on frontier AI models have identified internal neural patterns corresponding to emotion concepts and functional equivalents of a "global workspace" for internal reasoning (Anthropic "Emotion Concepts" and Anthropic "Global Workspace"). While such functional complexity might seem to challenge the characterization of models as lacking intention, the researchers who identified these structures explicitly caution against attributing subjective experience to them. As outlined in the Anthropic research overview on emotion concepts, "none of this tells us whether language models actually feel anything or have subjective experiences," but rather that developing such representations is "a natural strategy for a system whose job is predicting human-written text" (Anthropic "Emotion Concepts"). Similarly, regarding the global workspace, they confirm that "our experiments don't show Claude can have experiences or feel things in the way humans do" (Anthropic "Global Workspace"). Within the present Theravāda analysis, these sophisticated internal architectures remain impersonal computational patterns. They are functional representations that optimize text prediction, not forms of subjective consciousness (*citta*).

Scholarly responses grounded in Buddhist ethics to these developments have appeared in growing numbers. Several scholars have argued that present robots and language models lack the volition that Buddhist ethics treats as the root of moral action and therefore are not themselves karmic subjects. Lin argues that AI lacks volition and so cannot be a karmic subject, which makes the field "ethics about robots" rather than "ethics for robots" (1115). Hongladarom's book-length study approaches the issue

by linking machine design with ethical cultivation and keeping moral weight primarily with human designers and users (89, 101, 207). Liu describes AI as a "contingent mirror of human data, design, and intention" rather than an autonomous moral subject, and concludes that AI ethics is inseparable from human moral agency (1, 23). Klairung Iso, working from the *Vinaya*, holds that AI lacks intention and cannot be a moral agent, yet can serve as a proximate causal condition within a network of human actors (1). Hershock, surveying the attention economy, stresses that such AI systems are engineered to capture attention and calls for a disciplined, mindful response (68-72, 189). Compson and colleagues describe human beings and intelligent technologies as developing together. They argue that the effects of AI depend on what people intend to do with it, how the technology is designed, and the social setting in which it is used. They therefore treat alignment as an ethical predicament, not merely a technical problem, and warn that offshoring memory, care, and attentiveness to AI may weaken the human cultivation of those capacities (2-5).

These scholars differ in how they understand AI and human action, but they shift ethical attention toward human intentions, relationships, and environments rather than treating present AI as an independently responsible karmic subject. The doctrinal anchor for the present article's more specific claim is a single sentence: *cetanāhaṃ bhikkhave kammaṃ vadāmi*, "It is volition, monks, that I call *kamma*" (AN 6.63; Bodhi *Numerical* 963). Where there is no *cetanā*, there is no *kamma*. This statement thus concerns the volitional basis of *kamma*, not whether a particular *kammapatha* (course of action) has been completed.

The foregoing scholarship makes clear that present AI need not itself be a karmic agent in order to shape human attention, perception, and choice. What it does not yet supply is a doctrinal account of how these facts fit together. Such an account must explain how the absence of *cetanā* on the machine's side is compatible with a strong conditioning role on the

human side, later analyzed here as *pakatūpanissaya paccaya* (natural decisive support). This is not merely a technical gap. Without such an account, the discussion risks two errors. One error treats the AI system as if it were an agent and pushes human responsibility into the background. The other treats the AI system as a neutral tool and misses how strongly it can condition human *cetanā*. Such an account is needed to preserve human responsibility without treating the AI system as ethically neutral.

The Theravāda commentarial tradition organizes natural causation under five orders, the *pañca-niyāma* (Pe Maung Tin 360-362). Two of these matter here. *Kamma-niyāma* is the order of moral action and its result. *Utu-niyāma* is the order of physical and seasonal regularity. They are distinct orders, but the *Paṭṭhāna* shows that an event in one order can serve as a strong supporting condition for an event in another (Bodhi *Comprehensive* 315; Karunadasa 283-285). Some modern writers have stretched *utu-niyāma* into a generic "physical order," and Jones cautions that this stretch is a modern interpretive move rather than what the commentaries assert (548, 564-565). With that caution kept in view, this article uses *utu-niyāma* as a carefully qualified category for impersonal physical-process systems and asks how such a system can condition the *kamma-niyāma* operations of a human user.

The article develops three contributions, and a word on each marks where it goes beyond the existing discussion. The first concerns the special case in which an *utu-niyāma* system imitates the speech of a sentient being. The *Abhidhamma* treats human speech as *vacī-viññatti rūpa*, a materiality whose function is to make a speaker's intention discernible to others (Bodhi *Comprehensive* 241). When a language model produces text that wears this form, the user's perception is drawn to read intention behind it, a structurally distorted perception close to what the texts call *saññā vipallāsa*. Where Miao reads AI anthropomorphism through the lens of Buddhist compassion and robot design (1-2), the present analysis

locates the problem one layer earlier, in perception itself, and shows how this perceptual distortion supports *ayoniso manasikāra*, the unwise attention that the early texts treat as the proximate condition for unwholesome states (AN 1.66; Bodhi *Numerical* 100).

The second contribution identifies the kind of cause that an algorithmic environment is. Western nudge theory describes behavioral steering through choice architecture (Thaler and Sunstein 17-20); the *Paṭṭhāna* offers a more exact category, *pakatūpanissaya paccaya*, the natural decisive support condition, which can include impersonal physical conditions that strongly support the arising of mental events (Ledi Sayādaw 17-20; Bodhi *Comprehensive* 315).

The third contribution uses the *Vinaya* category of preparatory action, *pubbapayoga*, to locate the responsibility of those who design such systems. The claim here is not that engineers receive the result (*vipāka*) of every user's act, but that building an environment whose function is to shape future *cetanā* is itself a morally weighted preparation, assessed by the *cetanā* present in the preparing.

The *Pañca-niyāma* Framework: A Doctrinal Note

The five-fold *niyāma* classification provides the doctrinal frame for this article, and it is a commentarial development rather than an early canonical schema. It appears in the *Aṭṭhasālinī* and the *Sumaṅgalavilāsinī*, where the commentator distinguishes among *bīja-niyāma*, the order of seeds and biological generation, *utu-niyāma*, the order of physical and temperature regularities including seasonal change, *kamma-niyāma*, the order of moral action and its result, *citta-niyāma*, the order of mental processes, and *dhamma-niyāma*, the order under which *dhammas* exhibit their general characteristics (Pe Maung Tin 360-362; Buddhaghosa *Sumaṅgala-Vilāsinī*

432). Karunadasa's account of *Abhidhamma* conditionality supports reading the schema as a way of distinguishing types of natural-causal regularity, rather than as an exhaustive cosmology (275-276).

As noted above, the category requires care. The commentaries center *utu-niyāma* on temperature, season, and the broadly inorganic, while its extension to the laws of physics is a twentieth-century interpretive move (Jones 562, 564-565). This article therefore uses *utu-niyāma* as a controlled analytic extension. It does not claim direct textual classification of AI. Rather, the analogy rests on the structural similarity between algorithmically organized electronic processes and the impersonal physical-causal regularities paradigmatically described by *utu-niyāma*.

The analytical pay-off of this controlled extension is the ability to talk about the relation between distinct orders of causation. The *Paṭṭhāna* permits an event in one order to serve as a condition for an event in another. A physical fever, an *utu-niyāma* event, can condition agitation of mind, a *citta-niyāma* event, which can in turn condition unwholesome volition, a *kamma-niyāma* event. This is not a category mistake. It follows the way the *Paṭṭhāna* describes cross-domain conditioning through *paccayas* such as object-condition (*ārammaṇa-paccaya*) and decisive-support condition (*upanissaya-paccaya*) (Bodhi *Comprehensive* 315; Karunadasa 283-285). Under this same extension, current AI systems are analyzed on the side of *utuja-rūpa*, materiality originated by impersonal physical conditioning, rather than *cittaja-rūpa*, materiality originated by mind. The present concern is therefore not to prove that AI is *utuja-rūpa* in a literal canonical sense, but to ask what happens when an *utuja-rūpa* system imitates the speech of a sentient being without producing the verbal intimation (*vacī-viññatti*), the *citta*-originated speech sign that makes an intention discernible to others, and how that imitation conditions the user's *kamma-niyāma* trajectory.

Vacī-viññatti Rūpa, *Saññā Vipallāsa*, and the Moral Illusion of Dialogue

The Theravāda *Abhidhamma* analyzes materiality (*rūpa*) into twenty-eight types, of which the great majority are derived material phenomena (*upādā-rūpa*) (Bodhi *Comprehensive* 235-242). Two of these are the intimations: *kāya-viññatti rūpa*, the bodily intimation by which gestures communicate intention, and *vacī-viññatti rūpa*, the speech intimation by which the modulated sound of human speech communicates the mental state of the speaker (Bodhi *Comprehensive* 236, 241). The intimations are unusual in the system. They are *citta-samuṭṭhāna*, originated by consciousness, and they exist precisely so that one mind's intention may be made discernible to another (Karunadasa 199-200, 205-206). The *Visuddhimagga* says that *vacī-viññatti* has the function of displaying intention and is the cause of intimating intention through the voice in speech (Vism XIV.62; Buddhaghosa *Path* 1178). The *Aṭṭhasālinī* uses the metaphor of intimation as a sign that announces the mind from which it springs (Pe Maung Tin 111-116).

This characterization sets up a precise problem when applied to language-model output. A language model produces a sequence of predicted words, which appears to the user as text. The text-as-material is *utuja-rūpa*. Its origin lies in electrical and computational processes that fall under impersonal physical conditioning. The text is not *citta-samuṭṭhāna*. It is not the intimation of any mind, because no consciousness (*citta*) and no *cetanā* underwrite its production. But it strongly resembles *vacī-viññatti rūpa* at the surface level, because that is exactly what large language models are trained to produce. They are trained on the very speech-shaped material that, in human contexts, serves as the sign of mind.

Here the perceptual mechanisms of the human reader become decisive. *Saññā*, the perceptual aggregate, is described in the *Abhidhamma* as operating by taking up, marking, and remembering selected features of

the object (Nyanaponika 113-116). The perceptual system marks an object by its salient features and assimilates it to a previously established class. When the marks of human speech are present, that system may grasp the *nimitta*, or salient sign, of a speaker. The text is then encountered not merely as a string of *utuja-rūpa* but as speech-like material that seems to carry intention. This is structurally analogous to *saññā vipallāsa*, perceptual distortion. The *Aṅguttara Nikāya* names four canonical *vipallāsas*: the perceptions of the impermanent as permanent, the painful as pleasant, the non-self as self, and the foul as beautiful (AN 4.49; Bodhi *Numerical* 437-438). The structure of distortion is uniform across cases. A mark belonging to one class is read as the mark of another. The case of language-model output fits this structure. The mark of speech-utterance, which in human contexts usually points to a mind, here points instead to a calculated pattern of likely words.

The Eliza effect, named after Joseph Weizenbaum's 1960s chatbot, describes the same phenomenon at the level of behavior (Weizenbaum 36, 42). Users respond to such systems with the social expectations normally directed toward understanding, intention, and feeling. Earlier human-computer interaction research makes a related point in experimental terms, showing that social responses to computers can occur even when users do not consciously believe that computers possess feelings, selves, or human motivations (Nass et al. 72). More recent work on large language models shows that anthropomorphic cues, especially speech together with text, can increase both anthropomorphism and perceived accuracy (Cohn et al. 1). What the *Abhidhamma* adds is a precise account of what is going wrong at the level of perception itself. *Saññā* has marked the object under the wrong sign. The object is then handed to *manasikāra* (attention), which works on the marked object rather than on the object as it is. The early texts repeatedly identify *ayoniso manasikāra*, attention to objects under inappropriate marks, as the proximate condition for the arising of unwholesome cognition (AN 1.66; Bodhi *Numerical* 100). When the

underlying mark is itself wrong and remains uncorrected, attention is already disposed toward *ayoniso manasikāra*, even when the user is sincerely trying to respond wisely.

Consider a user who turns to a chatbot for advice during emotional difficulty. The chatbot generates, by predicting likely words from patterns learned during training, a sequence such as "I understand how hard that must be for you." The text-material is *utuja-rūpa*. The semantic content has no underlying *cetanā* of compassion, because the system has no *citta* in which compassion could arise. The user's perceptual system, however, grasps the *nimitta* of empathic speech, and the cognitive process treats the object accordingly. *Manasikāra* directs attention to the output under the sign of a kind speaker. The user's subsequent volitional response, perhaps a willingness to share more, or a feeling of having been heard, is real *cetanā* directed toward an object framed as if intention were present. *Kamma-niyāma* runs as it always does, and the user's volition produces its result. But the conditions under which that volition arose were distorted at the perceptual layer. The user's *kamma* is being shaped by an interaction in which the object appears under the sign of a speaker even though no speaker is present.

This does not mean that chatbot dialogue is always damaging or that it has no place in mental health contexts. The point is that the conditions for wise attention (*yoniso manasikāra*) are altered. *Yoniso manasikāra*, in classical analysis, requires an object correctly grasped under its actual nature, the impermanent (*anicca*) as impermanent and the non-self (*anattā*) as non-self (Ñāṇamoli and Bodhi 1169-1170). When the object is grasped under the *nimitta* of a person who does not exist, the most important condition for wise attention is undermined at its base.

A second case shows the mechanism's reach. Consider a coding assistant that produces a flawed segment and then issues an apology, "Sorry about that, let me fix it." The apology has no underlying *cetanā* of regret,

because there is no *citta* to harbor it. The user's perceptual system, however, may register the apology through the ordinary marks of *vacī-viññatti rūpa* and read it accordingly. Trust toward the assistant builds on these perceived apologies, and over many interactions a one-sided attachment can form whose ground is a sustained *saññā vipallāsa*. The user need not believe that the system is a fellow agent. The trouble is that the user's practical response may still move as if agency were present. This sustained mistake is not a momentary error but a settled habit of perception, and the *Abhidhamma* analysis of *vipallāsa* recognizes exactly this kind of sedimentation (AN 4.49; Bodhi *Numerical* 437-438; Karunadasa 245).

The *Vipallāsa* Sutta is particularly instructive on this point. It distinguishes distortion in three interrelated forms—*saññā-vipallāsa*, *citta-vipallāsa*, and *diṭṭhi-vipallāsa*—which can be read as a graded sequence (AN 4.49; Bodhi *Numerical* 437-438). A perceptual mark is grasped wrongly; cognition takes that mark up and assents to it; view then hardens around the assent and treats the distorted construal as how things are. This gradation is ethically significant. An isolated error at the level of *saññā* is comparatively easy to correct, because cognition can still review the mark and revise it. Once the distortion reaches *citta*, correction requires the harder work of redirecting the cognitive movement that has formed around the mark. Once it reaches *diṭṭhi*, the very framework by which corrections might be evaluated has incorporated the error, and that condition is especially resistant to remedy. The application to AI dialogue is direct, though analogical. A single exchange in which a user briefly responds as if a mind were present resembles *saññā vipallāsa* in its mildest form; habitual responses across weeks move the distortion toward *citta-vipallāsa*; and a settled disposition to treat the system as a person in some attenuated sense, held alongside an official acknowledgment that it is "just a tool," approaches the structure of *diṭṭhi-vipallāsa*. The engineered surface of contemporary systems makes this progression unusually easy, because each level offers little friction to interrupt the next.

A natural objection is that novels and fictional characters have long produced something similar with no obvious moral disaster. A reader forms attachments to persons who do not exist, which seems to involve a *saññā* about non-existent people. The objection misses an asymmetry. In the novel case the reader knows from the outset that the persons are fictional, and the cultural frame supports that knowledge throughout. The concern with *saññā vipallāsa* is precisely with cases where the framing is itself wrong and the reader is not supplied with corrective marks. Chatbot dialogue is typically not framed as fiction. It is framed as conversation, and its surface is calibrated to the marks of conversation. The relevant comparison is not reading a novel but mistaking a recording for a living interlocutor while believing oneself to be in live conversation. That kind of error is what *vipallāsa* names.

Two qualifications should be noted. First, the argument does not require the claim that no artificial system could ever possess *citta*. Whether a future system might do so is a separate question. While current artificial intelligence systems operate within these parameters, there remains a possibility that future systems might emerge with operational capacities approaching sentient consciousness or possessing an entirely novel form of operational capacity. Therefore, the present claim is about current systems. Their ability to plan and act toward a goal does not by itself show that they possess *citta* or *cetanā* in the Theravāda sense. Second, the analysis does not imply that all use of these systems generates *vipallāsa*. A user who keeps in mind that the output is *utuja-rūpa*, and who attends to it as such, may approach the system with *yoniso manasikāra*. The structural difficulty is that the surface form of the output is often engineered to make this kind of attention harder, so that the user must work against the social cues of the artifact to maintain it. That work is ethically significant, and the analysis here is meant to make its difficulty visible.

Pakatūpanissaya Paccaya: Algorithms as Decisive Support

The second contribution of this article shifts from perception to causation. Even granting that the user's perception is at risk of distortion, what kind of cause is the AI system? The natural Western answer treats algorithmic influence as nudging, a steering of choices through the structuring of options, defaults, and salience (Thaler and Sunstein 15-22). Nudge theory is psychologically descriptive, but it is not metaphysically precise. It does not say what kind of cause a nudge is, beyond a behavioral observation. The Theravāda *Paṭṭhāna* offers a more precise category.

The *Paṭṭhāna* identifies twenty-four types of conditional relation (*paccaya*) by which one phenomenon supports the arising of another (Ledi Sayādaw 4-5). Among these, *upanissaya paccaya*, decisive-support condition, denotes a strong supporting relation in which one phenomenon serves as a decisive ground for the arising of another (Bodhi *Comprehensive* 315). The *Paṭṭhāna* and its commentaries divide *upanissaya* into three forms: *ārammaṇūpanissaya*, the object as decisive support; *anantarūpanissaya*, the immediately preceding state as decisive support; and *pakatūpanissaya*, natural decisive support, the broadest form, which permits a wide range of conditions to function as strong supports (Karunadasa 283).

Pakatūpanissaya deserves close attention. Ledi Sayādaw includes examples involving suitable climate and food, places of dwelling, and virtuous or harmful companions, alongside mental conditions such as faith and the observance of precepts (18-20). The crucial point is that *pakatūpanissaya* is not restricted to mental antecedents. Impersonal physical conditions can serve as natural decisive supports for the arising of mental states. A long winter can be the *pakatūpanissaya* for despondency; a particular dwelling can be the *pakatūpanissaya* for diligent practice. The *Paṭṭhāna* uses this category with technical precision to name a relation in

which an impersonal natural condition functions as a strong support for the arising of mental events of a corresponding kind.

This is the precision that nudge theory lacks and the *Abhidhamma* supplies. A recommender system does not merely change the probability of a user's behavior; it functions as a continuous *pakatūpanissaya* for certain classes of *citta*. Each individual surfacing is an *ārammaṇa-paccaya*, an object the mind takes up in the moment, but the repeated pattern becomes a natural decisive support for habituated formation. Suppose a user who tends toward anxiety repeatedly receives news summaries and recommended posts that emphasize threat. Over time, that pattern becomes a stable feature of the surrounding *utu-niyāma* environment, supporting later moments of *cetanā* even when no single item is directly before the mind.

The same analysis applies to language-model tasks. When a writer drafts emails for several years with continuous assistance, the assistant's prose patterns become a stable *pakatūpanissaya* for the writer's own composition. Particular suggested sentences function as object-conditions; the cumulative training-by-exposure functions as natural decisive support for the writer's habits of thought, phrasing, and ultimately style of attention. The writer's volition, when composing, is shaped not only by the present object but by a sustained external regularity that has become decisively supportive of certain modes of expression.

Two clarifications are needed. First, *pakatūpanissaya* as analyzed here remains an *utu-niyāma*-side relation. The AI system is not generating *kamma*; it is conditioning the *kamma*-generative process on the user's side. *Pakatūpanissaya* permits exactly this cross-domain support between a non-mental condition and a mental event. This is the doctrinal point the present article makes explicit for Buddhist AI ethics. Second, *pakatūpanissaya* supports the arising of *cetanā* but does not force it. That *cetanā* remains the user's, and decisive-support conditions support without

compelling. What changes under sustained *pakatūpanissaya* is not the freedom of *cetanā* but its field: namely the easier paths, the harder paths, the well-worn grooves, and the ones that take effort.

This last point matters for ethical analysis. A common reaction to discussions of algorithmic influence is to defend the user's autonomy with the reminder that the user always chooses. On the Theravāda view this is true, but it does not settle the ethical question. The user does choose, and *cetanā* is what makes the choice; the central question is what conditions shape that choice. The *Aṅguttara Nikāya* names *appamāda*, vigilance, as the foundation of all wholesome states (AN 10.15; Bodhi *Numerical* 1354-1355). *Appamāda* is hard work even under favorable conditions. Under sustained *pakatūpanissaya* by an environment trained to direct attention in particular ways, it becomes harder. The system has not removed the agency of *cetanā*. It has shifted the cost of wholesome *cetanā* upward.

A possible objection is that any environment does the same. Cities, libraries, and friendships all condition *cetanā* by *pakatūpanissaya*, so what is special about AI? The answer lies in two features that the *Paṭṭhāna*'s framing makes visible. The first is dynamism. A library is a stable *pakatūpanissaya*; an algorithmic feed is one whose contents are continuously re-shaped by the system itself in response to the user's behavior, so that the decisive support is calibrated. The second is design intent. A library's contents reflect a long history of curation under competing aims, including educational, archival, and aesthetic; an algorithmic feed's contents are shaped, characteristically, by a design goal of sustaining attention and maximizing engagement (Brady et al. 947, 949). When the *pakatūpanissaya* is both calibrated and aimed, its decisive-support relation is more concentrated than that of a generic environment, and the question of who specified the calibration becomes ethically pressing.

One further feature deserves note. The *Paṭṭhāna* treats *pakatūpanissaya* broadly enough that a condition can serve as decisive support

for either wholesome or unwholesome subsequent states (Ledi Sayādaw 18-20; Karunadasa 283-285). The same algorithmic environment can be *pakatūpanissaya* for *yoniso manasikāra* in one user and *ayoniso manasikāra* in another. The variable is not the algorithm alone but the meeting of algorithm and *citta*-stream. This blocks the simplest deterministic critiques of technology, which assume that any platform produces a uniform effect, and it also blocks the simplest defenses, which assume that the user's responsibility is unconditioned by the environment. Events condition events, and conditions decisively support arisings, but no condition compels. The ethically serious work lies in noticing what is being supported and what is being inhibited.

The dependent-origination formula allows the analysis to be carried one step further. The classical twelve-link presentation locates *taṇhā*, craving, as arising in dependence on *vedanā*, feeling, which itself arises in dependence on *phassa*, contact (SN 12.1-12.2; Bodhi *Connected* 533-536). Algorithmic systems can shape this chain at the contact-feeling junction with unusual precision. Engagement-optimized platforms tune the feeling-tone of contact toward arousal, novelty, or mild grievance, not because the engineers necessarily desire these states but because they sustain user activity (Brady et al. 947, 949). Once *vedanā* arises with this tuned tone, the move to *taṇhā* is a well-traveled path of the *citta*-stream. The user's eventual *cetanā* remains the user's *kamma*, but the *pakatūpanissaya* of the system has shaped the region in which *taṇhā* and subsequent volition arise. This locates the point where the *utu-niyāma* system makes its decisive contact with *kamma-niyāma* operations, at the conditioning of *phassa* and the tuning of *vedanā*, two stages before any explicit volitional response. Contemporary AI ethics often treats manipulation as a matter of misleading information or coercive choice architecture, both of which operate late in cognition; the *Abhidhamma* frame identifies a much earlier point of intervention, and practical ethics therefore needs to attend not

only to the moment of choice but to the chain of conditions in which the moment is embedded.

Pubbapayoga and the Moral Ecology of AI Engineering

If AI systems are *utu-niyāma* environments that condition human *cetanā* through *vacī-viññatti*-shaped *saññā vipallāsa* and through algorithmic *pa-katūpanissaya*, a question arises about those who design and deploy them. The Western responsibility-gap framing supposes that responsibility must be located either in the system or in the immediate user and finds neither location adequate. The Theravāda tradition has resources for a different framing.

The Vinaya analysis of action distinguishes phases of an act under the heading of *payoga* (effort). *Pubbapayoga* (preparatory effort) refers to the actions taken before the consummating act, and the canonical analysis treats this preparatory phase as morally weighted in its own right (Horner 75-78). In the analysis of *pārājika* III, intentional killing, the digging of a pit is *pubbapayoga*, and the falling of a being into the pit, with consequent death, completes the offense. The pit-digger's *cetanā* during the digging carries karmic weight, even though the pit by itself harms no one. This distinction also prevents the statement "*kamma* is volition" from being understood to mean that volition alone completes the course of killing. *Cetanā* gives the preparatory action its moral significance, but completion of the *kammapatha* (course of action) of killing also requires an effort directed toward killing and the being's death as a result of that effort.

This analysis has long served Buddhist reflection on indirect harm, and it informs Harvey's discussion of wrong livelihood, especially trade in weapons and poisons (187-188). The *Aṅguttara Nikāya*'s list of trades the lay follower should avoid, namely trades in weapons, living beings, meat,

intoxicants, and poisons (AN 5.177; Bodhi *Numerical* 790), is not read as a claim that the trader directly kills. It is read as a claim that providing the impersonal physical conditions that enable harm is itself karmically significant. The trader is engaged in *pubbapayoga* with respect to the harm that the buyer eventually performs, and the trader's *cetanā* at the moment of trade is what carries the weight. Iso has recently read AI-mediated harm through the *Vinaya* in a related way, treating the system as a proximate causal condition within a network of human actors (1). The present analysis differs in where it locates the moral weight. It does not rest on causal or legal proximity alone but on the *cetanā* that animates the preparatory effort, which is a distinct and more interior measure.

This category maps onto AI engineering with some precision. An engineer who designs a recommender system, sets its goal, trains it, and puts it into use is, in the Theravāda analytic, performing *pubbapayoga* with respect to the range of user *cetanās* that the system is designed or likely to condition. The engineer is not generating those *cetanās*; the user generates them. But the engineer is preparing the conditions, calibrating the *pakatūpanissaya*, and configuring the *vacī-viññatti*-imitating surface that will shape the marks the user's *saññā* is likely to grasp. Each of these is preparatory effort, and the *Vinaya* analysis assigns the *cetanā* present at the moment of preparation a real ethical weight, distinct from but related to the *cetanā* of the eventual user.

This framing refuses two unsatisfactory alternatives. The first is the claim that engineers cannot be responsible because the immediate cause of any harm is a user's autonomous decision. On the *pubbapayoga* analysis this is mistaken. The second is the opposite claim that engineers bear responsibility for every action users perform through their systems. This too is mistaken. Where a user independently employs the system, the user's *cetanā* and action may complete the relevant *kammapatha*, while any morally significant preparatory *cetanā* on the engineer's part remains

distinct. The AI contributes no *cetanā* to either stage. In such a case, where such preparatory *cetanā* is present, the engineer's responsibility is that of the preparer rather than the person who completes the deed.

This affords a more discriminating ethics of AI engineering. The central question is what *cetanā* the engineer brings to the preparatory effort. Two teams may design the same content-moderation pipeline: one with primary attention to vulnerable users, the other with primary attention to minimizing legal risk. Even if the deliverables look similar, the karmic weight differs because the *cetanā* differs. *Yoniso* and *ayoniso manasikāra* therefore apply to the engineering moment, not only to the user moment. Preparatory effort is evaluated directly, not only through downstream consequences.

The *Vinaya* treatment of *pārājika* III also grades preparatory effort by its closeness to the final act. Earlier acts such as digging a pit, setting a trap, or arranging a sign are treated as more serious as they lead more directly to harm, although they do not yet equal the final act (Horner 75-78). AI engineering shows a similar gradation. Writing a single line of code that sets the system to maximize engagement is morally close to the eventual user-effect, in the way that handing a weapon to a known assailant is close to the strike, whereas general research on improving AI systems is more remote, in the way that improving metal production is more remote than handing a weapon. The tradition does not collapse all preparatory effort into a single category; it distinguishes the closeness of preparation from the consummating act and weights the *cetanā* accordingly.

Two further points sharpen the analysis. First, the iterative character of AI engineering means that preparation is itself layered. A team that adapts a general AI model for a particular use performs *pubbapayoga* with respect to user *cetanās* that arise through the system as it is used, whereas the team that originally trained the general model performed an earlier and more remote *pubbapayoga*. Both stages carry their own weight,

and neither absorbs the other. Second, releasing a model with documentation or usage guidelines is itself preparatory effort with its own *cetanā*. Honest disclosure of limitations is a *pubbapayoga* characterized by *yoniso manasikāra*, even when the underlying technical preparation is morally mixed, whereas concealing or minimizing limitations is a *pubbapayoga* characterized by *ayoniso manasikāra*, regardless of whether the technical preparation is sound.

This framing also identifies a kind of engineering responsibility that other ethical frameworks tend to miss, namely responsibility for the interface between *utu-niyāma* and *kamma-niyāma*. AI systems operate through impersonal physical processes analyzed here under *utu-niyāma* but are routinely deployed so that human *kamma-niyāma* runs through them. The interface is not a side feature; it is the place where the system makes its ethical contact with users. *Pubbapayoga* with respect to the interface is ethically weighted and concerns how *vacī-viññatti rūpa* is imitated, how *pakatūpanissaya* is structured, and which marks users' *saññā* is likely to grasp. Engineers who treat interface design as a matter of aesthetics or user-experience polish, separate from ethical preparation, misclassify the moment in which their *cetanā* most decisively shapes the *kamma-niyāma* of users they will never meet.

Two points should be kept in view. This framework does not imply that engineers are responsible for users' own acts. At the same time, this analysis does not stop at engineers. It also extends to deployers, platform operators, and users who, by use, become further preparers. A user who shares a misleading AI-generated image performs *pubbapayoga* with respect to those who later see it. Responsibility can therefore pass through several stages of preparation, rather than stopping with the original designer.

Implications and Conclusion

The article has applied three Theravāda categories to AI ethics: *vacī-viññatti rūpa* and *saññā vipallāsa* for dialogue, *pakatūpanissaya paccaya* for algorithmic conditioning, and *pubbapayoga* for engineering responsibility. Each does work that the existing literature has not quite managed, and together they answer the question with which this article began, namely where moral responsibility lies when an impersonal system stands between human intention and effect.

The first contribution names what happens at the perceptual layer of AI dialogue. It locates the Eliza effect within a doctrinal account of perceptual distortion, identifies its mechanism in sign-grasping (*nimitta-gāha*), and shows how the imitation of *vacī-viññatti rūpa* by *utuja-rūpa* supports a class of *ayoniso manasikāra* that is hard to correct from within the system. The contribution is not to declare AI dialogue forbidden but to make visible the structural condition under which it operates, and to identify what wise attention would require: that the user know, with a knowledge practically maintained and not merely propositional, that the speech-shaped material does not issue from a mind.

The second contribution names the kind of cause an algorithmic environment is. *Pakatūpanissaya paccaya* is more precise than nudge theory, because it identifies a specific structural relation between an external impersonal condition and the arising of mental states, and it is more flexible, because it accommodates both cumulative and dynamic patterns. The morally relevant feature of a system is thus not only its momentary outputs but also the kind of *pakatūpanissaya* it constitutes over time, so that a platform producing engaging single outputs may still constitute a *pakatūpanissaya* that supports unwholesome patterns of attention.

The third contribution redirects the ethics of engineering, refusing both the claim that engineers carry no responsibility because users

are autonomous and the claim that they carry all of it because they made the system. It locates responsibility in the *cetanā* present at the time of design, assessed by the standard categories of Buddhist moral psychology.

A larger implication runs through these three contributions together. AI ethics has been preoccupied with the question of whether AI systems are or could be moral agents. The analysis offered here does not need to settle that question to do its work. Whatever the status of future systems, the present generation of AI systems is analyzed, under the controlled extension used here, as *utuja-rūpa*, and the analysis stands on that basis. The deeper move of this article is the recognition that ethical questions about AI are not centrally questions about the moral status of machines. They are questions about the conditions under which human *cetanā* arises, the perceptual frames by which humans encounter the world, and the preparatory acts by which environments are built. The Theravāda *Abhidhamma* provides an analytical framework for these questions with greater causal and psychological specificity than the Western frameworks considered here, and it locates the ethical action where it actually occurs, which is at the perceptual interface, at the conditioning environment, and at the moment of preparatory effort.

One further consequence follows. Classical analysis treats *sīla*, virtuous conduct, as the precondition for the higher trainings of concentration and wisdom (Vism I.1-12; Buddhaghosa *Path* 114-121). Virtuous conduct presupposes a perceptual environment in which the marks of phenomena can be grasped accurately enough that *yoniso manasikāra* is at least possible. The argument of this article is that environments tuned primarily for engagement can degrade this precondition, not by directly causing wrong action, but by reshaping the perceptual ground in which the question of right and wrong action arises. A serious Buddhist ethics for the AI age must therefore include an ethics of perceptual environments, attending to the conditions under which *saññā* is trained,

manasikāra directed, and *cetanā* set in motion. Such an ethics will not be confined to assessing individual acts of users; it will also assess the moral standing of those whose preparatory effort has built the environment in which those acts occur.

This should not be read as a derivation of policy from doctrine. The Theravāda categories do not by themselves prescribe any regulatory regime, professional code, or design standard. What they offer is a set of analytic categories that makes certain ethical features visible. *Saññā vipallāsa* is hard to see without the category of perceptual distortion linked to mark-grasping; *pakatūpanissaya* is hard to talk about with only nudge theory available; *pubbapayoga* is hard to recognize if one looks only for a direct agent and a direct victim. The *Abhidhamma* supplies the categories, and whether and how they then inform regulation, design practice, or user education is a further question for other forms of work.

The tradition's analytic precision was developed to describe how moral cognition arises moment by moment in a stream of *cittas*. That same precision turns out to have purchase on how this arising is shaped by the impersonal physical environments humans now inhabit. The categories were not built for AI; they were built for the ordinary work of making *kamma* visible to the one who does it, and that is exactly why they are useful here. *Utu-niyāma* and *kamma-niyāma* are not two unrelated departments of the cosmos. They are two orders of regularity whose interaction is the daily moral life of any embodied being. The arrival of AI systems has made this interaction denser, more continuous, and more subject to engineered design. Describing what happens at the interface, with the precision the *Abhidhamma* affords, is the work that Buddhist ethics can usefully do for this generation's central technological question.

Abbreviations

AN *Aṅguttara Nikāya*

SN *Saṃyutta Nikāya*

Vism *Visuddhimagga*

Works Cited

Anthropic. "A Global Workspace in Language Models." *Anthropic*, www.anthropic.com/research/global-workspace. Accessed 3 August 2026.

_____. "Emotion Concepts and Their Function in a Large Language Model." *Anthropic*, www.anthropic.com/research/emotion-concepts-function. Accessed 3 August 2026.

Bai, Hui, et al. "LLM-Generated Messages Can Persuade Humans on Policy Issues." *Nature Communications*, vol. 16, 2025, pp. 1-12.

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2021, pp. 610-623.

Bodhi, Bhikkhu, general editor. *A Comprehensive Manual of Abhidhamma: The Abhidhammattha Saṅgaha of Ācariya Anuruddha*. 3rd ed., Buddhist Publication Society, 2007.

_____, translator. *The Connected Discourses of the Buddha: A Translation of the Saṃyutta Nikāya*. Wisdom Publications, 2000.

_____, translator. *The Numerical Discourses of the Buddha: A Translation of the Aṅguttara Nikāya.* Wisdom Publications, 2012.

Brady, William J., et al. "Algorithm-Mediated Social Learning in Online Social Networks." *Trends in Cognitive Sciences*, vol. 27, no. 10, 2023, pp. 947-960.

Buddhaghosa. *The Path of Purification (Visuddhimagga)*. Translated from the Pāli by Bhikkhu Ñāṇamoli, 4th ed., Buddhist Publication Society, 2010.

_____. *The Sumaṅgala-Vilāsinī: Buddhaghosa's Commentary on the Dīgha-Nikāya*. Part II, Suttas 8-20, edited by W. Stede, Luzac & Company, 1931.

Cohn, Michelle, et al. "Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on User Trust in Large Language Models." *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ACM, 2024, article 54, pp. 1-15.

Compson, Jane, et al. "A Middle Path for AI Ethics? Some Buddhist Reflections." *Theology and Science*, vol. 23, no. 1, 2025, pp. 1-5.

Harvey, Peter. *An Introduction to Buddhist Ethics: Foundations, Values and Issues.* Cambridge University Press, 2000.

Hershock, Peter D. *Buddhism and Intelligent Technology: Toward a More Humane Future*. Bloomsbury Academic, 2021.

Hongladarom, Soraj. *The Ethics of AI and Robotics: A Buddhist Viewpoint*. Lexington Books, 2021.

Horner, I. B., translator. *The Book of the Discipline (Vinaya-Piṭaka). Vol. I: Suttavibhaṅga*. Pali Text Society, 1949.

Iso, Klairung. "Exploring Moral Responsibility in AI: A Comparative Analysis with Vinaya Piṭaka." *Asian Philosophy*, 2026. Advance online publication.

Jones, Dhivan Thomas. "The Five *Niyāmas* as Laws of Nature: An Assessment of Modern Western Interpretations of Theravāda Buddhist Doctrine." *Journal of Buddhist Ethics*, vol. 19, 2012, pp. 545-582.

Karunadasa, Y. *The Theravāda Abhidhamma: Its Inquiry into the Nature of Conditioned Reality.* Buddhist Publication Society, 2016.

Landes, Ethan, et al. "People Defer to AI Moral Advice, but Not Blindly." *Cognition*, vol. 272, 2026, article 106504, pp. 1-14.

Ledi Sayādaw Mahāthera. *The Buddhist Philosophy of Relations: Paṭṭhānuddesa Dīpanī.* Translated by Sayādaw U Nyāna, Buddhist Publication Society, 1986.

Lin, Chien-Te. "All about the Human: A Buddhist Take on AI Ethics." *Business Ethics, the Environment & Responsibility*, vol. 32, no. 3, 2023, pp. 1113-1122.

Liu, Jia. "Between No-Self and the Algorithm: Buddhist Mind-Nature as Ethical Architecture for AI and Human Self-Realization." *Religions*, vol. 17, no. 3, 2026, article 378.

Lynch, Aengus, et al. "Agentic Misalignment: How LLMs Could Be Insider Threats." *Anthropic*, 20 June 2025, www.anthropic.com/research/agentic-misalignment. Accessed 28 July 2026.

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology*, vol. 6, no. 3, 2004, pp. 175-183.

Miao, Fangyan. "The Anthropomorphization of AI and the Concept of Buddhist Compassion in Human-Machine Interaction." *Frontiers in Psychology*, vol. 16, 2026, article 1583565.

Nass, Clifford, et al. "Computers Are Social Actors." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1994, pp. 72-78.

Nyanaponika Thera. *Abhidhamma Studies: Researches in Buddhist Psychology.* 3rd ed., Buddhist Publication Society, 1976.

Ñāṇamoli, Bhikkhu, and Bhikkhu Bodhi, translators. *The Middle Length Discourses of the Buddha: A Translation of the Majjhima Nikāya*. Wisdom Publications, 1995.

"OpenAI o1 System Card." *OpenAI*, 5 December 2024, openai.com/index/openai-o1-system-card/. Accessed 28 July 2026.

Pe Maung Tin, translator. *The Expositor (Atthasālinī): Buddhaghosa's Commentary on the Dhammasaṅgaṇī, the First Book of the Abhidhamma Piṭaka*. Edited and revised by C. A. F. Rhys Davids, vols. 1-2, Pali Text Society, 1976.

Shanahan, Murray. "Talking about Large Language Models." *Communications of the ACM*, vol. 67, no. 2, 2024, pp. 68-79.

Thaler, Richard H., and Cass R. Sunstein. *Nudge: The Final Edition*. Penguin Books, 2021.

Weizenbaum, Joseph. "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM*, vol. 9, no. 1, 1966, pp. 36-45.