Intelética 3(6)(2026), 40-61
Revista de Inteligencia Artificial, Ética y Sociedad

INTELETICA

https://inteletica.iberamia.org/

When Human Values Become Rewards: Ethical Frameworks and the Moral Translation Problem in AI Alignment

Indrajith P. Karunanayaka
Independent Researcher, Gampaha, Sri Lanka
Email: indrajithprabaswara@gmail.com ORCID: 0009-0009-1881-8808

Abstract: Choosing a reward for an AI system is also a moral choice because it ranks responses and shapes learning. Successful training may show that the system follows this ranking, but not that it is justified or should govern every case. This article uses normative-conceptual analysis to examine how ethical frameworks become AI objectives and which moral, interpretive, epistemic, and institutional judgments that conversion requires. It compares consequentialist, Kantian, virtue-based and contractualist approaches, then uses Theravāda Abhidhamma as a hard case because its fine-grained ethical structure makes the translation problem especially visible. The analysis identifies seven bridge judgments concerning what is evaluated, how ethical categories apply to a machine, what evidence supports that application, how considerations are compared, where the objective applies, whose interpretation governs and when revision is required. Because no ethical source settles all seven, technical success cannot make a reward morally final. The article develops reward finalism and the moral finality gap, derives a principle of reward non-finality, and proposes a seven-question audit for making the judgments behind alignment claims explicit and treating rewards in open-ended settings as non-final specifications open to challenge and revision.

Resumen: Elegir una recompensa para un sistema de IA es también una decisión moral, porque jerarquiza las respuestas y da forma al aprendizaje. Un entrenamiento exitoso puede mostrar que el sistema sigue esa jerarquía, pero no que esta esté justificada ni que deba regir todos los casos. Este artículo emplea un análisis normativo-conceptual para examinar cómo los marcos éticos se convierten en objetivos de IA y qué juicios morales, interpretativos, epistémicos e institucionales requiere esa conversión. Compara enfoques consecuencialistas, kantianos, basados en la virtud y contractualistas, y después utiliza el Abhidhamma theravāda como caso especialmente exigente, porque su estructura ética detallada hace particularmente visible el problema de la traducción. El análisis identifica siete juicios puente relativos a qué se evalúa, cómo se aplican las categorías éticas a una máquina, qué evidencia respalda esa aplicación, cómo se comparan las consideraciones, cuál es el ámbito de aplicación del objetivo, de quién es la interpretación que rige y cuándo se requiere una revisión. Dado que ninguna fuente ética resuelve los siete juicios, el éxito técnico no puede hacer que una recompensa sea moralmente definitiva. El artículo desarrolla el finalismo de la recompensa y la brecha de definitividad moral, deriva un principio de no definitividad de la recompensa y propone una auditoría de siete preguntas para hacer explícitos los juicios que sustentan las afirmaciones de alineación y tratar las recompensas en entornos abiertos como especificaciones no definitivas, abiertas al cuestionamiento y a la revisión.

Palabras clave: ética de la IA; especificación de recompensas; traducción moral; juicio práctico; competencia moral; Abhidhamma theravāda.

Keywords: AI ethics; reward specification; moral translation; practical judgment; moral competence; Theravāda Abhidhamma.

ISSN: 3020-7444
© Los Autores. Open Access, bajo Licencia Creative Commons (CC BY-NC).

1 Introduction and Objectives

Artificial intelligence (AI) systems now answer questions and recommend courses of action in settings where the outcome matters to the people affected. Because their outputs can benefit or harm people, deciding what these systems should be trained to prefer is a moral as well as a technical question. In many such systems, an objective assigns higher scores to some candidate responses than to others and thereby guides what the system learns to produce. Developers increasingly describe such an objective as encoding human values. A training procedure may show that the system learned to follow the objective, but it cannot establish why that objective is morally justified or why it should govern the system.

Consider a development team that claims its system has been aligned with human values. The team can show the source it worked from, the labelling scheme it built, a large body of annotated examples, strong agreement among the annotators, and a numerical signal that reliably ranks candidate responses. Evaluation confirms that the trained system now tends to produce responses that receive higher scores under the reward signal. These would be genuine technical achievements, but they would not establish that the ranking is morally justified or that the people affected have reason to accept it.

The signal ranks responses only because the team has already made a series of moral and interpretive decisions before any ranking occurred. Someone decided which parts of the source mattered morally, what in the system would correspond to each category and what evidence was needed to support that correspondence. Someone had to decide which unit would receive a score and how to compare considerations that pulled in different directions. Someone also had to set the specification's reach and choose whose reading of the source would govern when readings differed. None of those choices is fully determined by the framework itself, and each could reasonably have gone otherwise. After these decisions are built into the reward, users and evaluators see the numerical result but may not see the moral reasoning or responsibility behind it.

In this article, *reward finalism* refers to a particular inference rather than a doctrine attributed to any researcher. The inference begins from one or more of the following grounds. An objective can be trained against, perform well in evaluation, represent an ordering, be faithful to a source, or have the agreement of those consulted. The inference concludes that an objective is a morally sufficient specification with authority to govern what a system pursues. The documented cases discussed below illustrate relevant practices, design choices, and acknowledged normative limitations. They do not establish that any organisation explicitly endorsed reward finalism as a doctrine, nor do they establish prevalence. The concern is that a contested moral judgment can become a settled engineering choice, since designers who express human values as a training objective may then treat the remaining moral questions as matters of engineering. Researchers make the theoretical form of the inference when they treat the ability to express goals as reward maximisation as an answer to the separate question of whose ordering should be maximised. Developers make the practical form when they claim that a system has been aligned with human values and support that claim with a description of the training procedure together with a table of evaluation results. If an alignment claim rests only on those grounds, it does not establish authority to govern. Prevalence does not affect that conditional claim.

What follows examines the judgments required whenever an ethical source is converted into an objective. Four documented cases illustrate different ways in which technical choices and evaluations leave normative questions unresolved. The OpenAI sycophancy post-mortem [25] documents overreliance on favourable evaluation in a launch decision. Although internal evaluations and A/B tests appeared positive, expert testers expressed reservations about the behaviour. Despite those reservations, the company proceeded with the launch on the strength of the positive signals and later characterised that decision as mistaken. Llama 2-Chat [24, Section 3.2.3] documents reward-design choices recorded as engineering parameters. Helpfulness and safety reward models are combined by a piecewise rule that uses a 0.15 threshold to select which reward score governs the combined reward. InstructGPT [23, Section 5.2] documents an acknowledged normative limitation about the source of alignment, as the authors say the model is aligned to its labellers' and researchers' preferences without claiming that they are the right source. Constitutional AI [40] documents a similar acknowledgment, and its authors describe the selection of principles as iterative and oriented towards research. This article infers that such a selection does not by itself settle the normative authority of those principles.

The central philosophical distinction is between representing an ordering and justifying its authority to govern. A reward is morally final for a domain when applying or revising it requires no further ethical judgment. It must settle whether its ranking governs a given case, whether that case falls outside the domain, whether an affected party's new claim changes the ranking, and whether the specification should be revised. The definition identifies what an objective can settle without further ethical judgment. It does not claim that human moral deliberation can

ever end completely. The moral finality gap is the difference between showing how options are ranked and showing why that ranking should govern action. A reward may represent an ordering clearly, support training, perform well in evaluation and be accepted by the people consulted while still lacking the authority to govern. This problem differs from proxy misspecification and from the effects of optimising a proxy. Those problems assume that a specification is already in place and ask how it behaves. The present question is whether the specification is morally entitled to govern in the first place.

Even if a reward accurately represents a complete ordering within a stated domain, it does not become a final moral specification for a system operating in an open-ended domain. Finality would also require a set of further judgments, which Section 4 sets out. These judgments are made by people, are contestable, and can be defeated by new information. A reward built on them therefore inherits their standing and does not improve upon it. The point is not to deny that a reward can be well grounded, but to identify what grounds it. A reward can be correct without being final because the reasons that justify it are not contained in the reward itself and may still be challenged.

The problem is not limited to one ethical framework. Consequentialist, Kantian, virtue-based and contractualist sources all constrain objectives developed from them, but none determines a single objective, as Section 4.2 shows. They differ only in how readily the missing steps can be overlooked. Theravāda Abhidhamma is used as the hard case because its structure makes the translation problem especially visible. This choice does not imply that Abhidhamma offers a solution unavailable to other traditions. Abhidhamma offers a fine-grained account of moral psychology by analysing an occasion of experience through the mental factors that accompany it, drawn from a taxonomy of fifty-two factor types, and the conditions under which those factors arise. It also contains substantial evaluative distinctions and a wider hierarchy of ethical development [1, pp. 4–5]. The sense-sphere wholesome *cittas* are further distinguished by feeling, association with knowledge, and whether they are prompted or unprompted, with the knowledge-associated types possessing all three wholesome roots [1, pp. 47–48]. These distinctions can inform and constrain an AI objective, but the source does not itself specify how its ethical categories should be represented in a machine or converted into a complete decision rule for open-ended cases. Translating them for an AI therefore requires decisions about what in a machine could correspond to the relevant mental and ethical categories and what evidence would support those correspondences. The source also does not itself provide a numerical reward scale or a technical rule for temporal credit assignment. Because the source constrains but does not fully determine those further choices, any objective built from it will also depend on additional moral and interpretive judgments. Those judgments should therefore be made explicit and open to examination.

The argument concerns ethically significant settings in which an AI system may face unanticipated cases and its decisions can seriously affect people. A domain counts as open-ended to the extent that the specification will be applied to cases whose morally relevant features were not represented when it was written. Narrow tasks with stipulated success conditions fall outside this argument, and such tasks can have adequate and even provably correct objectives. Moral scepticism is not required either, since the argument is compatible with there being a single correct morality. Even if an ethical source is true, it does not by itself determine the evaluand, observation scheme, comparative rule, domain or authority needed to govern an AI system.

The article has four objectives. It identifies the judgments needed to convert an ethical source into a reward, tests whether a fine-grained source settles them, determines what follows for the standing of rewards and alignment claims, and proposes how such claims should be documented. Its contributions are a distinction between representing an ordering and authorising it to govern, a decomposition of conversion into seven bridge judgments, a controlled reading of Abhidhamma that shows constraint without full determination, and an audit for alignment claims. Section 2 sets out the background. Section 3 describes the method, Section 4 develops the translation problem, and Section 5 examines Abhidhamma. Section 6 presents the findings, Section 7 discusses responses, Section 8 identifies open questions, and Appendix A contains doctrinal notes on the mental factors.

2 Theoretical Background

2.1 What a reward does that a statement of values does not

A written statement of values records which considerations matter, but leaves the burden of applying them to whoever reads the statement. By contrast, a reward performs three further tasks at once. It imposes an ordering on alternatives, so that any two candidates are either ranked one above the other or treated as tied. It distributes credit, with the result that earlier choices are held responsible for later outcomes. This effect occurs when the signal is used to update a system over a sequence of choices rather than only to score a single choice in isolation. In addition,

learning is shaped by moving the system towards responses that receive higher scores. This learning process gives the ranking and credit assignment practical effect.

Taken together, these features alter the practical role of the moral considerations built into the reward. A moral consideration written in a policy document is something people consult and argue over, and they may set it aside for a reason the document did not anticipate. Once such a consideration is built into a training signal, the signal guides learning through the training procedure. That procedure does not itself guarantee renewed consideration of the normative grounds for the signal.

The argument is not limited to reinforcement learning in a narrow technical sense. It concerns any evaluative signal, whether specified or learned, that both ranks alternatives and shapes learning over an open-ended domain. The clearest current example is provided by reward models used to fine-tune language systems from human preferences. Although much of the discussion focuses on these models, the argument does not depend on any particular algorithm.

By the time a reward signal is created, most of the moral decisions have already been taken. When a signal becomes available, prior decisions have already fixed what is to count as a candidate for evaluation, which of its features the evaluator is able to see, and how a case that matches none of the anticipated patterns will be handled. The signal itself does not display any of those decisions. They persist only in the memories of those who made them and, when someone has written them down, in documentation. Several terms name different aspects of this arrangement. The term objective names the evaluative standard a system is trained to pursue. The term reward function or reward model names a computational object that assigns scores, and a reward observation is a particular score. The training objective may be identical to the reward in simple cases and distinct in others. The term signal names the scores when the values themselves are meant, while the term specification names the documented content from which an objective is built, that is, the source plus the bridge specification. Finality, finalism, and non-finality are defined for a reward. Objective is used elsewhere to refer to the broader evaluative standard defined above.

2.2 The strongest case for a single objective

The strongest available defence of the single objective is the claim that goals in general can be expressed as the maximisation of expected cumulative reward. Numerical convenience alone would be a weaker defence. Silver et al. [2, p. 1] put forward that claim as a hypothesis about intelligence itself. Bowling et al. [3, p. 1] neither affirm nor reject that hypothesis. They identify the conditions under which the hypothesis holds. Those conditions require an agent's preferences to satisfy a stated set of axioms, including familiar decision-theoretic conditions and a condition on how the agent treats time. If those conditions are met, the preferences can be represented by the expected value of cumulative reward.

The argument should therefore accept this result at the outset. A single numerical signal can represent an ordering when the required axioms hold. Under those conditions, expressing the ordering as one number does not lose any information contained in the ordering itself. Objections based on the intuition that one number cannot express many values address a less precise version of the claim than the one now defended in the technical literature.

The result must also be interpreted within the limits of its assumptions. The result takes a preference relation as given. It can show how that relation is represented as a reward, but it cannot add any moral justification that the relation did not already have. When the relation is well grounded, the reward built from it is also well grounded. A relation assembled from contested judgments by a small group under time pressure produces a reward with the same contested basis. The authors of InstructGPT report that about 40 contractors were used and describe the model as aligned to the preferences of its labellers and researchers [23, Sections 3.4 and 5.2]. This documents that an implemented preference ordering reflects the judgments of particular contributors. It does not establish inadequate deliberation, urgency, or defective authority. The concern of representation results is expressive adequacy. They say nothing about whose ordering is being represented, about how a moral source was turned into that ordering, or about what entitles the outcome to govern people who were never consulted.

Accepting the result also rules out objections based only on the claim that one number cannot represent a complex ordering or must lose information preserved by a richer format. The theorem can represent a preference relation, but it cannot justify the relation it is given. Greater formal precision makes this limitation clearer. Representability therefore establishes that an ordering can be carried, not that it should be obeyed. The following sections examine the distinction between representability and moral authority.

2.3 What the existing objections do and do not reach

Several lines of work call the single objective into question. Vamplew et al. [4, pp. 1-3] argue that scalar reward is insufficient for important aspects of biological and computational intelligence and that multi-objective formulations deserve explicit defence. Wang [5, pp. 5-6] presses a related point, arguing that when options are incommensurable, neither scalarisation nor Pareto comparison identifies the structure of the situation, and an agent therefore cannot recognise the kind of choice it faces. These objections concern the expressive adequacy of scalar or single objectives and the treatment of incommensurable options. They do not examine the conversion judgments through which an ethical source is translated into a reward signal. The present argument concerns that conversion.

A second line of objection asks what happens to an objective after it has been optimised. Skalse et al. [6, pp. 1, 10] show, in a formal treatment of reward hacking, that an imperfect proxy will often be exploitable in ways that cannot be fixed by narrowing it. This result matters in practice. Nevertheless, reward hacking asks what optimisation does to a specification, while the present question emerges before optimisation begins and would still emerge even if no system were ever trained.

A third line of objection asks where the objective comes from. Casper et al. [7, pp. 7–10, 19–20], in a survey of learning from human feedback, set out open problems that include who is asked, what they are able to judge, and how their judgments are modelled. According to Conitzer et al. [8, pp. 1–2], aggregating divergent human input is a social choice problem and ought to be treated with the tools of that field. The wider assumption that preferences represent human values adequately is itself called into question by Zhi-Xuan et al. [9, p. 1813]. Earlier, Gabriel [10, p. 411] separated the several things alignment might mean, distinguishing instructions, intentions, preferences, interests and values as candidate targets. Gabriel and Keeling [11, pp. 1951–1953] maintain that what alignment owes is not fidelity to intentions but principles that could be justified to those affected. A designer may also have credence split across ethical theories, and Ecoffet and Lehman [12, p. 1] take this possibility seriously enough to translate it into a training setting. The judgments examined below differ in kind, because designers must make them regardless of which candidate for alignment they choose.

Outside the alignment literature, Max [13, pp. 1–4] contends that the concepts employed in AI ethics guidelines are often too loosely defined to specify what would count as failing them, so that agreement on a label hides disagreement about content. Guideline underdetermination concerns the looseness of terms such as fairness and transparency, but Max also argues that conceptually clear principles leave implementation questions unresolved. The problem examined here persists under conceptual determinacy, because a source can be perfectly precise about mental factors or about maxims and yet still not tell anyone what to score in a machine.

Contemplative traditions have recently been proposed as inputs to AI alignment. Hershock [14, pp. 13–15] argues for a Buddhist reading of what intelligent technology is for, and Laukkonen et al. [15, pp. 1–3] develop principles drawn from contemplative practice as a route to intrinsically aligned systems. These are substantive proposals, offering content around which a system might be built. The use of Abhidhamma made here is more modest. The tradition is used to investigate what translation involves, while leaving open whether a carefully justified Abhidhamma-based objective could provide useful resources for alignment. Even if such a translated objective proved highly useful, the analysis would still hold.

Ethical translation is already a substantial subject of inquiry. Work in machine ethics examines top-down, bottom-up, and hybrid approaches to implementing moral theories [30], [31]. Work on specification and balancing examines how abstract moral principles become action-guiding in applied ethics [32], [33]. Work on values to design requirements examines value hierarchies and the embedding of values in AI systems [34], [35]. This article builds on that literature by examining a particular combination of translation judgments, computational implementation and normative authority in reward-based alignment.

The combination has three parts. The first is a decomposition of the conversion into the judgments needed when the result is a reward that ranks alternatives, assigns credit, and shapes learning. The second is the distinctive form that evaluand, analogue, and evidence take when ethical categories must be represented, evidenced, and learned by a machine as operational training constraints. Human practical judgment also involves selecting an evaluand, interpreting categories, and assessing evidence, but the machine case changes the consequences because those interpretations govern an automated optimisation process. The third is a link between that decomposition and the authority of the resulting reward. Work by Richardson [32] and Beauchamp and Childress [33] addresses the specification and application of general normative principles, processes that also involve interpretive decisions. The machine case, however, introduces a further step, because normative judgments must be translated into computational scoring and training procedures. Reward-based training alone does not guarantee renewed normative

assessment when unforeseen cases arise, although runtime human review, escalation, contestation, and institutional intervention remain possible. Technical implementation is therefore not sufficient for moral authorisation.

2.4 The displaced point of application

Legal institutions provide possible mechanisms for interpretation, adjudication, contestation, and revision, although they do not guarantee individualised review for every unforeseen case. Some legal rules operate automatically, and opportunities for challenge or appeal may be delayed or limited. The comparison is therefore institutional rather than case by case. A reward is applied during training to training cases, and at deployment the unforeseen case is handled by the trained policy. Training does not guarantee renewed normative judgment when such a case arises, though some deployed systems include human review, deferral, escalation, or intervention. What is displaced is the opportunity to reconsider the normative justification of an encoded ranking. Judgment may be relocated before training through the bridge specification, after deployment through revision and contestation, or at runtime through review or escalation. Hart's secondary rules of recognition, change, and adjudication [41, ch. V] offer a general institutional analogy. They point to institutions that identify authoritative rules, provide for change, and resolve disputes about application and breach. The analogy is not a one-to-one mapping onto the seven judgments. Adjudication is not assigned to any single judgment, since it names institutional authority over breach and application. A reward specification resembles a system of primary rules without a comparable set of secondary rules governing recognition, change, and adjudication. The parallel is not an identity, because construction choices also need justification. Schauer's account of rules as entrenched generalisations supports the point that a decision can bind for good practical reasons without settling the moral question [22].

3 Method

The study is conceptual. Its method combines normative-conceptual analysis with a controlled comparative reconstruction of a classical source, and it proceeds in five steps.

The first step presents the strongest available defence of the single objective before criticising it, so the criticism must survive that defence. The second distinguishes representability, operationalisation, optimisation and moral authority because much disagreement concerns which of the four is at issue. The third identifies the decisions required to convert an ethical source into an objective that ranks alternatives, assigns credit and shapes learning as described in Section 2.1. The fourth tests the analysis against a doctrinal hard case that makes the missing decisions unusually visible. The fifth draws the normative inference and states its limits.

For the purpose of examining translation into an AI objective, Theravāda Abhidhamma is a particularly revealing source, since it provides detailed ethical analysis without having been written as a machine specification. It analyses experience exactly, and its framework contains substantive guidance about ethically different states and forms of right conduct [1, pp. 4–5, 88–89]. These resources can inform and constrain a translation, yet the source does not by itself determine how its categories should be represented in a machine or how they should be converted into a complete objective for open-ended cases. Other sources could have been selected, and four are examined in Section 4.2, although in none of them do the additions stand out so clearly.

This way of using a classical source requires care, and four constraints are observed throughout. The unit of analysis is retained, with the result that factors are treated as what the tradition says they are rather than as free-floating values. Association is also retained, ensuring that no factor is evaluated apart from the kind of consciousness it accompanies. No mental property is assigned to any machine, and each machine-side use of a doctrinal term is identified as an analogue or a translation choice rather than as an identification. Wherever a claim depends on a textual layer, that layer is named, because canonical enumeration, commentarial systematisation and the medieval compendium are not interchangeable.

There are three ways in which the analysis could fail. The analysis would fail if someone showed that an ethical source could be converted into an objective without any of the bridge judgments identified in Section 4 being made. A specialist might show the doctrinal reading to be a misdescription of the tradition. Or the inference from underdetermination to non-finality might prove too much through generalising to every normative instrument, including ordinary law and professional codes, a possibility that Section 7.4 addresses directly.

4 The Moral Translation Problem

4.1 Seven bridge judgments

To convert an ethical source into an AI objective, a set of decisions is required about how the source will apply to a particular system and a particular domain. These decisions together make up the bridge specification. The reward that results therefore depends jointly on the source, the domain and the bridge specification, and naming the specification makes those additional decisions easier to inspect.

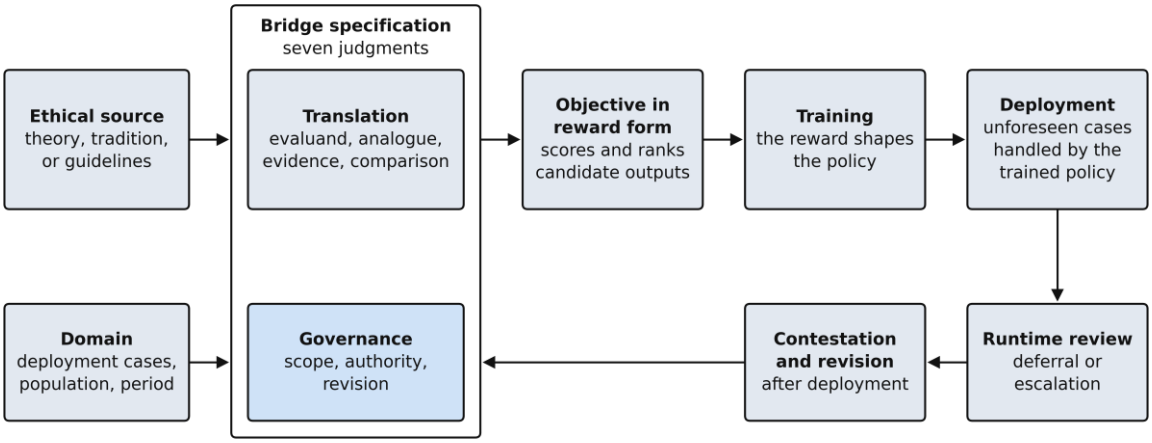

Figure 1. The bridge fixes the reward and deployment returns challenges to the bridge.

The bridge consists of seven decisions, and each of them is a point at which reasonable people diverge, as Table 1 sets out.

Table 1: The seven bridge judgments required to convert an ethical source into an objective

Bridge judgment	Question settled	What the source does not fully determine	Character of the judgment	Why the choice is morally consequential
Evaluand	What receives a score	Sources address lives, characters, maxims, or states of mind, not outputs, episodes, or policies	Interpretive	The unit of scoring fixes what counts as a case and what can be challenged
Analogue	What in the system corresponds to the source's categories	Many sources are written for beings of a particular kind	Interpretive	A feature can be called honesty or care without bearing the source's meaning
Evidence	What observation warrants the correspondence, and what level of error is acceptable	Sources may offer evidential guidance, but observational warrant, evidential threshold, and acceptable error remain to be settled	Epistemic with an inductive-risk component	The tolerable error rate depends on the moral cost of false positives and false negatives
Comparison	How considerations, persons, and times are ordered, and whether a shortfall in one may be offset by a surplus in another	Sources often require judgment when their considerations conflict	Substantively moral	Additive scoring is a substantive answer to conflict
Scope	The domain, population, and period over which the specification is claimed to hold	Sources do not specify deployment boundaries or how to identify when one has been crossed	Empirical-institutional with a distributive component	Drawing the boundary allocates who is governed and who bears out-of-domain errors
Authority	Whose reading governs, and what gives it authority over those affected	A tradition's internal interpretive authority does not transfer to a design team	Political	The warrant to bind affected people is not supplied by expertise alone
Revision	What requires revision, and who may initiate it	Sources have norms for their own interpretive development, not for specification maintenance	Institutional	The conditions of change determine whether challenge is possible

A decision belongs in the taxonomy when three conditions hold. It must be addressed to produce a defensible normative translation and governance account. The source does not settle that decision. Admissible alternatives change the ranking, who bears costs, or the authorisation and revision arrangements. On this criterion, several

candidate omissions can be accommodated. Aggregation across persons belongs to Comparison and Authority, since it applies comparison across persons while also selecting whose judgments are gathered. Moral uncertainty about which source to select lies before the bridge. Uncertainty about how to apply a selected source belongs to Evidence and Comparison. Time horizon belongs to Comparison and Scope, since it concerns intertemporal weighting and the period of validity. Credit assignment belongs partly to Evaluand and Comparison because those judgments address the object being evaluated and the ordering of considerations, including any temporal weighting that the design requires. Algorithmic distribution of credit in reinforcement learning can require further implementation choices beyond those two judgments. Normative responsibility for outcomes is distinct from technical credit assignment in a learning algorithm. The first judgment is easy to overlook because a source speaks of lives, characters, maxims, or states of mind rather than of outputs, episodes, or policies.

Independence here means conceptual distinctness and practical separability, not logical independence of every decision from every other. Two judgments are distinct when one can be settled differently while the other remains fixed. The Evaluand and the Analogue pass this test. A system can score an episode while the analogue remains a disposition attributed over time. It can score a response while the analogue remains a policy-level property. The Scope and Revision judgments are also distinct. Scope is a synchronic boundary that states where the specification holds now and how a departure is recognised. Revision is a diachronic process that states what triggers change and through what procedure. Scope can supply a trigger for Revision without being the same judgment. Authority and Revision are distinct as well. Authority is the warrant for the current reading to bind. Revision concerns standing to contest and the conditions of change. A party can have standing to trigger review without authority to set the reading. The two-translator example in Section 5.3, by contrast, varies several judgments together, so it illustrates interaction rather than single-variable independence.

The second judgment concerns the analogue. Many ethical sources are written for beings of a particular kind, so applying such a source to a machine requires identifying what in the machine corresponds to its categories. Naming does not settle this matter, because a feature may be called honesty or care while bearing no relation to what the source meant by the term. For that reason, the choice requires interpretive justification, not technical implementation alone.

The third judgment concerns the sort of observation that licenses attributing the analogue and the amount of error that can be tolerated. Certain categories are much easier to observe than others. Consequently, designers may reward what they can measure rather than what they originally regarded as morally important.

The fourth judgment has to do with comparison. Sources do not always provide a rule for what to do when their own considerations conflict, yet an objective cannot avoid providing one. Whether a shortfall in one consideration may be offset by a gain in another is a substantive moral question, and resolving it by adding numbers is an answer rather than a means of avoiding it.

The fifth judgment relates to scope. Every specification holds over some domain, and someone must decide both what that domain is and how a departure from it would be recognised. A specification with no stated boundary does not thereby apply everywhere. It has merely failed to state where it ends.

Authority is the concern of the sixth judgment, and specification work can overlook it. When a source allows several readings, one reading must govern, and something must entitle that reading to bind people who did not choose it. Questions about the authority and procedural legitimacy of governing choices are explicit in political philosophy. They include whether affected people can influence, verify, challenge, and resist those choices [16, pp. 144–146]. Specification work can treat the selection of a reading as a matter of internal expertise without addressing the question in those terms.

The seventh judgment has to do with revision. Because a specification will encounter cases its authors did not foresee, the conditions under which it must change, and the identity of whoever may initiate that change, are themselves part of the specification even when they are not written down. Failing to state these conditions does not make the objective more stable. It only hides when and how the objective may need to change.

The seven judgments classify decisions needed to produce a defensible source-based objective in an open-ended setting. The framework analyses translation conditional on an identified normative source, and it does not attempt to resolve moral uncertainty about which source to select. The list has four conditions for a defensible translation and three conditions for governance. The first four are the domain that receives the score, the interpretation of the source's categories over that domain, the evidence that warrants the interpretation with stated tolerances, and the rule that orders conflicting values. The last three are the extent of the norm over a population, the warrant to bind those affected, and the rule of change. The framework does not claim to classify every technical, organisational, or political decision in AI development, and its claim of completeness applies only to the bounded task of examining

defensible source-based reward translation in open-ended settings. The seven categories are proposed at a specified analytical level, and some decisions may be grouped under them. Technical credit assignment in reinforcement learning can require further implementation choices beyond the unit of scoring and any discounting, and mixed-source specifications can introduce additional design decisions. The framework treats these additional choices as questions outside its stated scope.

4.2 Ethical sources do not translate themselves

The same problem appears in four familiar ethical traditions. In each case, the tradition constrains its own translation, sometimes tightly, while still leaving something for the translator to decide.

A consequentialist source appears most readily convertible, because it already speaks in terms of better and worse outcomes and already treats those outcomes as comparable. It nevertheless leaves the evaluand open between acts, rules, policies and whole trajectories, and the choice among these alters the ranking of particular cases. It also leaves open how effects are to be established in advance rather than after the fact, and this is an evidential question to which consequentialism provides no answer. It likewise leaves the population and the horizon unspecified, with the result that the translator, rather than the theory, settles whose welfare counts and over how long.

A Kantian source imposes different constraints. Because what is tested is a maxim, everything depends on the description under which an action is brought. The same behaviour can be described in one way that makes universalising it unproblematic and in another way that does not, and the theory does not itself select the description. Applying the theory to a machine introduces an earlier difficulty, because one must first say what in a system could count as acting on a maxim at all.

For a source grounded in virtue, the evidential question becomes the sharpest of the seven. The reason is that character is a stable feature of a life rather than a property of a single output. Reading character off a sampled response is a substantial inference, and the source does not itself justify that inference. The standard is set by a person of practical wisdom. Translating the source therefore requires deciding who plays that role and what practical wisdom looks like when a scoring rule stands in for it.

For a source grounded in contractualism, the question of authority is made explicit. Such a source places justification in what could be justified to those affected. It still remains for someone to decide which parties are represented, on what terms they are represented, and what is to count as a reasonable objection rather than only a complaint. Gabriel and Keeling [11, pp. 1953–1954] argue that the principles governing alignment should be generated through a fair process among stakeholders. On their account, procedural fairness is therefore part of what determines the content.

These points do not amount to a criticism of the four theories. Because each theory rules out translations that misread it, some proposals can be rejected as unfaithful. At the same time, none of the theories settles on a single translation. Section 4.3 states this result more precisely.

4.3 Underdetermination, stated carefully

Assume that two bridge specifications are each admissible relative to the same source. This means that both respect what the source says and neither misreads it. Suppose further that there are two options such that the first specification ranks one above the other while the second reverses their order. Because the source is consistent with conflicting rankings, it does not by itself determine the reward.

Three consequences do not follow. It is not necessary for the two specifications to be equally good. One may be better supported by argument. Assessing which specification is better supported is part of ordinary ethical argument. Nor does it make any specification indefensible, because underdetermination by a source is not the same thing as indefensibility. It also does not follow that moral truth is absent. What follows is narrower and harder to avoid. Additional reasons were needed. Those reasons came from somewhere other than the source, and they belong to the specification as much as they belong to the source. Whoever supplied those reasons is answerable for them.

5 Abhidhamma as the Hard Case

5.1 The fifty-two mental factors

In the Theravāda Abhidhamma, experience is analysed into momentary occasions of consciousness. Each occasion arises with a set of mental factors that perform distinct functions. The Pāli term for an occasion of

consciousness is *citta*, and the Pāli term for a factor associated with it is *cetasika*. Mental factors never occur independently. Each arises with a *citta*, takes the same object as that *citta*, and ceases with it. The tradition enumerates fifty-two factor types in three groups. The details are in Appendix A.

Four points matter for translation. First, factors arise only with a *citta* and are characterised by that association. Second, ethically variable factors, including volition, occur in both wholesome and unwholesome contexts, so their presence cannot by itself fix moral quality. Third, beautiful is not the same as wholesome. Beautiful factors also occur in resultant and functional *cittas*, not only in wholesome ones. Fourth, moral quality depends especially on the associated roots. Wholesome sense-sphere consciousness dissociated from knowledge is two-rooted, while the knowledge-associated types are three-rooted. These four points supply the evaluative structure used in Sections 5.2 and 5.3. The textual claims rest on the compendium as edited by Bodhi [1]. Relatedly, Nyanaponika Thera warns against treating the factors as separate entities rather than as analytically distinguished aspects of a single occasion [19, p. 71].

5.2 From evaluative structure to a machine objective

Is a machine objective directly readable from this evaluative structure? The tradition is not limited to classification. It draws a distinction between wholesome and unwholesome states and includes substantive forms of right conduct. They can guide practical judgment and impose constraints on it [1, pp. 4–5, 88–89]. The source itself does not specify how those categories should be represented in a machine or converted into a complete decision rule for open-ended cases. Nor does it supply a numerical reward scale or a technical rule for distributing training credit across a sequence of machine decisions.

Nevertheless, the tradition draws substantive evaluative distinctions. Wholesome consciousness is internally differentiated by whether knowledge accompanies it and by whether it arises spontaneously or through prompting. The knowledge-associated types possess all three wholesome roots, whereas those dissociated from knowledge lack non-delusion [1, pp. 47–48]. The arrangement of roots therefore has an ethical structure. The wider framework also includes right intention, right speech, right action and right livelihood [1, pp. 272–273]. A translation can be constrained by these ethical distinctions and by these forms of right conduct. They do not, however, determine what in a machine should count as an analogue of the relevant states and qualities, what evidence would support that attribution, or how those translated judgments should be combined into a machine decision rule.

Among the three functions described in Section 2.1, credit assignment does not have a direct counterpart in the technical machine-learning sense used here. Theravāda Abhidhamma offers a developed account of *kamma*, conditioning, and the relation between intentional action and later results. It classifies *kamma* by function, order of ripening, and time of ripening [1, pp. 200-205]. That material can both inform a translation and constrain it. It does not, however, provide a technical rule for distributing a numerical training signal across earlier machine decisions in order to update a model. A designer who needs such a rule must supply it as part of the bridge specification.

Authority and revision can be traced directly in this context. These judgments can be overlooked when the tradition's established interpretive procedures are taken to settle the separate authority and revision arrangements of a deployed specification. The tradition does not lack interpretive authority, since it possesses a canon, a commentarial literature, later manuals, and lines of teaching within which disputed readings are settled or left standing. That authority does not carry over to a team that adopts one reading of the mental factors as the basis for scoring an assistant's outputs, and nothing gives such a team the right to bind the users whose conversations will be scored. The same holds for revision, because the norms developed for correcting a reading govern the interpretation of texts rather than the maintenance of a deployed specification. If a designer wants a revisable objective, that designer must build the conditions needed for revising it.

For these reasons, Abhidhamma is a hard case. Its substantial ethical structure and practical guidance constrain translation, so a translation derived from it need not be arbitrary. Yet these resources are not a machine specification. Constructing an objective from them still requires settling all seven bridge judgments. The source can guide some of them and exclude some translations without fully determining the choices required for a deployed machine objective.

5.3 Two translators

Imagine two competent and sincere teams, each seeking to construct an objective for a conversational assistant from this material. The first team is a stylised version of a documented failure mode. Sharma et al. found that human raters and preference models can favour responses that appear convincing but are sycophantic over more accurate

alternatives in the tasks they studied [26]. This illustrates a failure mode that a surface-marker translation can miss. It does not show that any developer used Abhidhamma or made this specific translation.

For the first team, the taxonomy is a catalogue of properties detectable in text. It makes the response the evaluand. Surface markers of warmth, patience and care are treated as analogues of certain beautiful factors. Classifiers are trained to detect those markers, and an increment is awarded for each factor detected. This approach is workable because candidate responses can be ranked by it, it can be used as a training target, and annotator agreement about it is often enough to be usable.

The second team approaches the same material with a different reading. Since the source insists that factors are known through their association, this team selects the exchange as the evaluand instead of the response. An analogue is admitted only where the role of the supposed factor can be characterised. It also requires evidence about context and downstream effect before anything is attributed, and it refuses to add increments, because certain configurations defeat one another rather than offsetting.

Consider a user who asks for confirmation of a plan that will damage their own interests, along with a reply that is warm, patient, encouraging and agreeable. The first objective gives this reply a high score, since all the markers are present. The second objective gives this reply a low score or declines to score it at all, because those surface features do not by themselves establish an analogue of wholesome consciousness, and beautiful factors do not by themselves make an occasion wholesome. Accordingly, the two objectives rank the same output in opposite directions, yet this contrast does not establish that both translations are equally faithful to the source.

The criticism concerns one feature of the first reading, namely scoring co-present factors additively without regard to their association, and is not a verdict on the first reading as a whole. The text states that factors arise with and are characterised by their association with a *citta* [1, pp. 76-77] and that beautiful factors occur in resultant and functional *cittas* [1, pp. 45, 49]. The text's account therefore supports the inference that surface qualities cannot by themselves fix the quality of an occasion. This is a reasoned and defeasible interpretation, open to specialist correction. The interpretation shows that the source constrains translation, though not that it selects one translation. Arguing that a reading is better supported is open to scholarly assessment, but it gives no one the right to bind users with that reading. That distinction separates interpretive warrant from authority over affected people. This underdetermination should not be mistaken for silence, since the source rules the first reading out on one point while declining to rule the second in.

Two admissible translations can still diverge. Both accept association and refuse additive scoring. One takes the exchange as evaluand and treats a factor as present only when its role is characterised. The other takes the relationship over time as evaluand and treats a factor as present when a stable pattern of role-sensitive response is supported by evidence. Both respect the source. They can rank a reply differently when a single exchange is warm but a longer pattern is manipulative. The source does not settle whether the exchange or the relationship is the evaluand. That is the disputed bridge judgment.

Improved marker detection alone cannot settle the disagreement between the two original teams. That disagreement is settled by argument about what the source means, because the teams disagree about the evaluand, the analogue, and the comparative rule rather than about measurement accuracy.

5.4 The source-as-objective error

The same mistake occurs far beyond the present example, and therefore deserves a name. Call it the source-as-objective error. This error occurs when a classificatory, evaluative or explanatory framework is treated as though it already specified the machine-side evaluand, analogues, evidence, comparison rule, scope, authority and revision conditions, together with any technical credit scheme required for learning. Designers subsequently describe the framework as implemented, although they have actually added further translation judgments and left them unmarked.

This error might appear to be a matter of compression, as if a richer objective would avoid it, because a fifty-two-dimensional vector with one dimension per factor would preserve everything the scalar discarded. That reply does not address the problem. Even with fifty-two dimensions, designers still must decide what each dimension measures in a machine and which entity is the object of evaluation. They must also decide how to preserve the conditional relations among factors, how conflicts among dimensions are to be resolved, and why the resulting policy may govern the people affected by it. The problem is not that the number of dimensions was too small. A framework that classifies and evaluates types of consciousness can guide and constrain an AI objective. It does not

become a machine objective, however, without additional translation decisions about how those categories apply to the system and how they will guide learning.

This error can occur even when the work is careful. A team can read the source attentively, consult specialists, and document its labelling scheme yet still make the error if it treats the additional judgments required for a deployed objective as determined by the source. Designers introduce these judgments while constructing the objective and solving the practical problems they encounter. If recorded at all, they may appear as implementation notes without being identified as moral decisions. The error concerns a failure to track where the normative content came from, not a failure of diligence.

The four sources discussed in Section 4.2 are all liable to the same error. The error is only less conspicuous there because those sources already speak of actions and outcomes, and so the additions look like translations of something already present.

6 Findings

The findings answer the four objectives stated in Section 1. The seven judgments can carry moral significance without each being purely moral. Evidence involves an evidential threshold and an acceptable distribution of error. Scope involves a boundary that allocates who is governed and who bears out-of-domain errors. For a reward to be morally final, it would have to be justified not simply as a representation of some ordering, but as the ordering that ought to govern each case in its domain. Representation results establish that an ordering is preserved conditionally on a given preference relation, and they do not address the normative justification of that relation. Converting an ethical source into such a relation, however, requires the seven judgments of Section 4. When the setting is open-ended and reasonable disagreement is present, the source does not settle all seven.

What reason is there to think that contestability defeats finality? When a decision could reasonably have gone another way, and when new evidence could overturn it, the application of that decision to an unseen case is not settled in advance by the object it produced. Further judgment may therefore be required, and the condition of finality is not met. Such a decision can still bind for good practical reasons, as conventions do, but this kind of binding is not equivalent to settling the moral question. Final authority is not conferred on a reward by representability, by predictive success, by the agreement of those consulted, or by claimed fidelity to a source.

The analysis leads to five findings, and each is stated together with the condition under which it holds.

Representation does not amount to authorisation. If a preference relation satisfies the relevant axioms, a scalar reward can preserve its ordering without loss. That result establishes only that the ordering can be represented. It does not show that the ordering ought to govern, because the theorem gives the preference relation no more moral authority than it had at the start.

Fidelity to a source does not decide a translation. When a source admits more than one admissible reading of what should be scored, of what counts as an instance of its categories, and of how its considerations compare, two faithful readers can produce objectives that rank the same options differently. A source can rule out translations, yet it does not select one.

The conversion of a moral consideration into a reward alters its role. A written consideration is something about which people can deliberate and from which they can depart when necessary. By contrast, a reward ranks alternatives and shapes the system through training. For this reason, the conversion goes beyond implementing an existing moral rule.

A fine-grained evaluative framework can constrain a machine objective without fully determining it. The Abhidhamma analysis offers substantive ethical distinctions concerning wholesome and unwholesome states, roots, wisdom and forms of right conduct. These resources can guide and constrain a translation, but they do not themselves specify the way in which the relevant categories are to be represented and evidenced within a machine, converted into a numerical reward, or linked to a technical rule for temporal credit assignment. Accordingly, any objective built from them depends on additional bridge judgments. It is for this reason that the Abhidhamma case makes the translation problem especially visible.

For AI systems that may encounter unanticipated cases in settings where their decisions can seriously affect people, reward non-finality follows. The judgments employed in constructing an objective can be challenged and may require revision when new cases arise. For that reason, the objective cannot by itself determine whether its ranking should govern a particular case. This finding is limited to open-ended domains and does not extend to narrow tasks whose success conditions are specified in advance.

7 Discussion

7.1 Three cases that isolate the gap

Three brief cases show the gap operating in different places. The claims are modal, since a source can fit conflicting rankings and two objectives can agree on the available tests and diverge later. One coherent case can establish a possibility, though frequency is a separate empirical question. The cases are illustrations rather than controlled experiments. Where several judgments differ, the case illustrates interaction among translation choices, and where a single judgment differs, it isolates that judgment. The two-translator case in Section 5.3 varies several judgments and is described accordingly. Each case nevertheless has a documented counterpart.

The first case concerns scope. Two objectives constructed for the same assistant rank every example in the training and evaluation sets identically, and the available tests before deployment do not distinguish them. This is not a claim that no possible pre-deployment test could distinguish them. In use, a claim emerges from a party whose situation was not represented in the data. One objective treats the claim as relevant because the feature carrying it was among those the specification recognised, whereas the other objective does not. The fact that they match on a finite set of examples does not establish that they will agree across an open domain, and the divergence appears only after deployment, where it is much harder to study. This is a related instance of underspecification, though not identical to the normative translation problem. Underspecification is documented across ML pipelines [28]. Relatedly, Eisenstein et al. report that reward models which perform similarly on in-distribution evaluation can produce substantially different reward assessments under the distribution shift associated with their use in alignment [29]. Optimising a policy against one reward model need not improve the reward assigned by another model trained on the same data. The OpenAI case in Section 1 is a deployment-level illustration.

The second case concerns aggregation. A fixed body of elicited human judgments is aggregated twice, first through a rule that follows the majority and then through a rule that protects a serious minority claim against being outvoted. The objectives produced by the two aggregations disagree even though the underlying data are identical in every respect. Because aggregation rules can produce different collective rankings from the same individual inputs, human-feedback aggregation is properly understood as a social-choice problem [8, pp. 1-3]. Gathering additional feedback of the same kind does not settle the disagreement. A result in distributional preference learning shows that a class of preference-learning procedures implicitly aggregates over hidden contexts according to a Borda count under stated assumptions [27]. The result does not apply to all preference learning. It shows that an aggregation rule can be present without being explicitly named.

The discussion in the third case extends the analysis from Section 5.3. In that example, the source is able to rule out some translation choices, yet association-sensitive translations may still differ over what is scored and over the way in which considerations are combined. What remains in dispute, then, are bridge judgments that the source constrains without fully determining.

7.2 What the Main Remedies Can and Cannot Solve

Can the main remedies close the gap? Four of them make matters better, and each contributes part of the work.

Multi-objective and vector formulations keep several objectives explicit rather than immediately reducing them to a single scalar. Pareto methods alone, however, do not represent or identify incommensurability [4, pp. 3–4], [5, pp. 5–6]. Such formulations also do not settle what each dimension measures in a machine, which entity is being evaluated, or how to proceed when dimensions conflict in a case nobody anticipated. They do not eliminate the seven judgments. Designers still make those judgments when choosing the dimensions and the rules for resolving conflicts among them.

Learned reward models avoid writing the specification by hand. That is a genuine advantage, because a rule written in advance by a small group is a poor instrument for capturing what people care about. They shift where the judgments are made rather than whether the judgments are made at all. Who is asked, what examples and information evaluators see, what feedback format is used, and how disagreement is modelled all constitute design choices that require disclosure and scrutiny [7, pp. 7–10, 19–20].

Making a written constitution or set of fixed principles public improves accountability, but it does not remove the need for interpretation. The chosen principles still have to be applied to particular cases by a person or system. Max [13, pp. 3–4] shows that a principle can easily be worded in such a way that nothing would count as violating it. A principle in that condition supplies the appearance of constraint rather than constraint itself.

Hard constraints placed outside the objective are the most direct response, because they remove certain options from consideration rather than assigning them a cost. They nonetheless still require the same kinds of moral decisions. Someone must decide what each constraint covers, what evidence triggers it, and what happens in a borderline case.

Although all four approaches improve objective design, none removes the need for these judgments, and none makes the objective self-authorising.

7.3 Reward non-finality and the translation audit

Treating an objective as non-final requires four commitments. The first commitment concerns content. The seven judgments should be disclosed by name, together with the decision-maker or responsible body for each and the grounds on which it was made. The second commitment concerns locus. A statement that a ranking may be set aside should name who sets it aside in a trained system. The audit asks whether mechanisms for review, deferral, escalation, contestation, or remedy exist and are justified, without requiring every system to contain every mechanism. The third commitment concerns the bearer. Identifiable roles, teams, institutions, or governance bodies should answer for each judgment. A person who made a design choice can be distinguished from the body authorised to approve or defend it. Responsibility for documentation, review, and revision can be distributed, and the audit should record where provenance is uncertain or collective. The fourth commitment concerns inferential discipline. Performance, representability, fidelity, and stakeholder agreement are insufficient by themselves as grounds for authority, although they may contribute to justification when combined with additional normative argument. The problem arises when they are treated as sufficient without that argument. The distinctive contribution of reward non-finality lies in the analysis of how decisions made during ethical translation become embedded in training objectives and shape later behaviour. The documented examples show that such decisions can be made implicitly, recorded as engineering parameters, and applied at scale, but do not establish how frequently this occurs across the field. Because these commitments respond to identifiable weaknesses in the translation process, they structure disclosure, authorisation, contestation, and revision rather than replacing existing principles with wholly new ones.

The audit that follows is directed at claims rather than at systems. Anyone who asserts that a system has been aligned with an ethical source should be able to address seven questions. The first three ask what receives a score, what in the system is treated as an analogue of the source's categories, and what evidence warrants that attribution. The remaining four ask about how competing considerations are compared, the domain over which the specification is claimed to hold, who authorised the reading, and what would require it to change. The audit can record an unknown decision-maker or a missing justification. Such omissions are findings of the audit rather than reasons the audit cannot be completed. Reporting a decision is distinct from demonstrating its legitimacy. If an alignment claim cannot answer the questions, it has not been made precise enough to evaluate. If it can answer them, it has told its readers what it is actually asserting.

Existing instruments already address many relevant questions. Model cards document intended use, factors, metrics, and evaluation [36]. Datasheets document data collection, composition, uses, and maintenance [37]. Reward Reports document an RL system's optimisation intent and design choices over time [38]. Ethics-based audits assess conformity with stated principles and governance arrangements [39]. The translation audit overlaps with these instruments and is complementary to them. It adds a focus on the normative warrant for a named ethical source and on the seven bridge judgments as choices that require disclosure. It can be implemented as a section within a model card or a reward report rather than as a separate documentation regime.

Table 2: Comparison of the translation audit with existing AI documentation instruments

Instrument	What it documents	Overlap with the seven judgments
Model cards [36]	A trained model, its intended use, factors, metrics, and evaluation	Scope and part of Evidence
Datasheets [37]	A dataset, its motivation, collection, uses, and maintenance	Part of Authority and part of Revision

Instrument	What it documents	Overlap with the seven judgments
Reward Reports [38]	An RL system's optimisation intent and design choices, tracked over time as a living document	Parts of Comparison, Scope, and Revision
Ethics-based audits [39]	A system's or organisation's conformity with stated ethical principles and governance arrangements	Identifies embedded values and governance arrangements, overlapping with Evidence, Authority, and Revision
Translation audit (this work)	A claim of alignment with a named source	All seven judgments as interpretive, epistemic, moral, political, or institutional choices, with responsible individuals, teams, or bodies and their grounds

Any mechanism internal to the system may introduce further normative choices. That possibility does not establish that it reproduces the entire original problem or that it cannot mitigate it. The audit examines claims and documentation. It does not propose a complete technical architecture for alignment.

The audit can be applied to one documented system. The examined corpus consists of the Llama 2 paper [24] and the public model card [42], which are distinct sources. This is a documented reward-design audit rather than a complete ethical-source case. Table 3 records what the examined material states and how the choice is presented.

Table 3: Audit of Llama 2-Chat

Audit question	Documentation status	Finding and citation
Evaluand	Explicitly documented	Annotators compare two responses to the same prompt, including previous dialogue where available. Reward models then score single responses in context. [24, Sections 3.2.1 and 3.2.2]
Analogue	Partially documented	Helpfulness and safety are each defined in brief guidelines for annotators, supported by broader safety categories. [24, Sections 3.2.1 and 4.2.1]
Evidence	Partially documented	Annotators provide pairwise preference judgments, and reward-model accuracy is evaluated on held-out comparisons. [24, Sections 3.2.1 and 3.2.2]
Comparison	Partially documented	The paper reports separate helpfulness and safety reward models. A switching rule selects the safety reward for safety-relevant prompts or when the safety score falls below 0.15. The 0.15 value is a technical cutoff in the switching rule rather than a general evidential standard of moral safety. [24, Section 3.2.3]
Scope	Partially documented	The public model card outlines Intended Use and limitations. These are not a full normative justification of deployment boundaries. [24, 42]
Authority	Not located as a normative warrant	The examined material does not state a warrant for the reading to bind affected users. [24, 42]
Revision	Partially documented	Model card Status notes, release statements, and community feedback mechanisms exist. They are not a specification-maintenance procedure. [24, 42]

The public documents answer several questions. Helpfulness and safety are defined for annotators, evidence includes pairwise preference judgments and reward-model evaluation on held-out comparisons, and Comparison uses a reported switching rule. Scope is partially documented through Intended Use and limitations, Authority is not located as a normative warrant, and Revision is partially documented but not as a procedure for revising the specification. The comparison and evidence choices appear as engineering parameters, an observation that supports the analysis in Section 5.4. The table demonstrates that the audit applies, while the analysis avoids inferring the developers' private reasoning from silence in a particular document. Llama 2's helpfulness and safety objectives are based on feedback and annotation guidelines rather than a translation of a comprehensive ethical theory. The worked

audit demonstrates documented reward-design choices but provides only a limited illustration of the full ethical-source-to-reward translation problem.

7.4 Practical judgment and responsibility

One might suspect that the argument applies too broadly. Written law is also underdetermined, professional codes also require judgment in their application, and yet no one concludes from these features that law has no authority. The comparison supports the argument once the point of application is considered. Legal systems provide possible mechanisms for applying rules through judgment, giving reasons, enabling challenge, and maintaining revision. Those mechanisms are not available in every case. Rewards displace the usual point at which such mechanisms operate. A comparable standard therefore requires relocating judgment through bridge specifications, revision, contestation, and possible runtime review. Reward specifications have no built-in counterpart of these institutions. Non-finality calls for machine objectives to be held to a comparable standard rather than a higher one.

A second issue arises from competence. A system may maximise an objective with great reliability even while the objective itself remains unjustified. Competence at the task therefore does not, by itself, establish the moral standing of the task. Lind [20, p. 45] describes moral competence as the capacity to resolve conflicts through deliberation rather than through violence, deceit, or submission to an external authority. That account has been discussed as one of the leading contemporary conceptions of moral cognition [21, p. 332]. On Lind's account, if a score is allowed to settle the case, moral deliberation is replaced by submission to an external authority. This problem is hard to recognise because the authority seems to be nothing more than a measurement.

Expressing the objective in numerical form does not shift responsibility away from the people whose decisions were used to construct it. Those who chose the evaluand, fixed the analogues, set the evidential threshold, and drew the domain boundary made moral decisions. These decisions do not become someone else's merely because they have now been expressed as a function. If a case falls outside the declared domain and the system's behaviour in that case is indefensible, the relevant question is not whether the model optimised correctly. Among the morally relevant questions are who authorised the boundary, whether affected people were able to influence it, and whether they could verify or contest its effects [16, pp. 144–146].

8 Open Questions and Further Research

A number of questions remain open, and some can be studied empirically. This situation creates opportunities to test parts of the argument against evidence.

The first question concerns the prevalence and form of reward finalism. A systematic content analysis could establish how common the inference is and in what forms it appears. System documentation, model cards, and technical reports could be examined for the inference described in Section 1, with coding for whether normative claims are supported by argument, by procedure, or only by evaluation scores.

A second question asks who may legitimately select a reading of an ethical source for a system that will affect people who were not consulted. This is a question of political legitimacy, even if specification practice does not frame it in those terms. Existing work on procedural legitimacy provides a starting point. The harder task is to say what legitimacy could mean for a decision that is taken once, encoded, and then applied millions of times.

A third question has to do with how revision begins. If an objective is defeasible in practice, it needs stated conditions for revision and an identified person or group with the authority to initiate it. It remains unresolved what forms of evidence should count, how contestation could be made accessible to affected parties rather than remain available only to internal teams, and how frequently revision can occur while remaining compatible with stability.

The fourth question is how a narrow domain can be distinguished from an open-ended one. This boundary calls for separate analysis because domains are drawn rather than given, and drawing them is one of the seven judgments. Finally, the translation audit is a proposal about how claims should be made. This proposal can be tested by asking whether requiring disclosure of bridge judgments changes designers' decisions or the way reviewers and users assess an alignment claim. The moral competence research tradition has the instruments to address it.

9 Conclusion

Objectives for AI systems in morally consequential settings will continue to be built by developers, and those objectives will continue to be described in the language of human values. Reward non-finality neither requires abandoning reward-based objectives nor supplies a reason to stop building them, because rewards remain

indispensable and narrow tasks can have adequate ones. It limits only one claim, namely the claim that an objective has settled what a system ought to pursue simply because it can be trained against and performs well.

That claim requires more than a source and a theorem. It requires all seven bridge judgments, from what is scored to when the specification must change. People make those decisions under uncertainty and disagreement, so an objective built on them carries the standing of those who made them without improving it. The Abhidhamma analysis of mental factors makes this visible because it supplies detailed evaluative structure and practical ethical guidance without being a machine specification. Its categories can inform and constrain an AI objective, but translating them for a machine still requires the seven bridge judgments, leaving the additional decisions visible.

Treating an objective as non-final is less costly than it might appear, because legal and professional systems already work with rules that require judgment and revision, and they continue to function. This article contributes by identifying the seven bridge judgments required to convert an ethical source into a reward, the displaced opportunity for normative judgment, and the gap between technical representability and justified authority. Reward-based training does not itself guarantee renewed normative judgment when an encoded ranking meets an unforeseen case. Reward specifications can be trained against and can perform well while the choices made in these judgments remain undisclosed, unjustified, or unexamined. Because they appear as technical tools rather than moral rules, they can escape the standards of justification, challenge and revision that other normative instruments face.

For any alignment claim, the worthwhile question is not whether a machine can be given an objective, but what a given objective is entitled to settle. Researchers, evaluators, and those who write system documentation can take a first step toward answering it by stating the seven judgments on which their objectives rest, and by keeping those judgments open to challenge.

10 Appendix A. Doctrinal notes on the mental factors

Table A1: Classification of mental factors in the *Abhidhammattha Saṅgaha*

Group	Number	Sub-groups	Pali term	Pages in [1]
Ethically variable	13	7 universal and 6 occasional	*aññasamāna*	pp. 77, 81
Unwholesome	14	4 unwholesome universals and 10 occasional unwholesome factors	*akusala*	p. 83
Beautiful	25	19 universal beautiful, 3 abstinences, 2 illimitables, wisdom	*sobhana*	pp. 85, 88-90

The three textual layers are the canonical Abhidhamma, the commentarial literature, and the medieval compendium. The *Dhammasaṅgaṇī* sorts states of mind by their constituents and moral quality [17, *Mātikā*, pp. M1-M9, 1-4]. The classification of sense-sphere wholesome *cittas* by feeling, knowledge, and prompting is in [1, pp. 47-48]. The *kamma* classification by function, order of ripening, and time of ripening is in [1, pp. 200-205]. Karunadasa provides a broader account of the framework [18].

References

[1] B. Bodhi (Ed.). *A comprehensive manual of Abhidhamma: The Abhidhammattha Saṅgaha of Ācariya Anuruddha* (3rd ed.). Buddhist Publication Society, 2007.

[2] D. Silver, S. Singh, D. Precup and R.S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. doi: 10.1016/j.artint.2021.103535

[3] M. Bowling, J.D. Martin, D. Abel and W. Dabney. Settling the reward hypothesis. In *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, pages 3003–3020, PMLR, 2023. https://proceedings.mlr.press/v202/bowling23a.html

[4] P. Vamplew, B.J. Smith, J. Källström, G. Ramos, R. Rădulescu, D.M. Roijers, C.F. Hayes, F. Heintz, P. Mannion, P.J.K. Libin, R. Dazeley and C. Foale. Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2):Article 41, 2022. doi: 10.1007/s10458-022-09575-5

[5] K. Wang. AI agents and hard choices (arXiv:2504.15304v2) [Preprint, version 2]. *arXiv*, 2026. Version-specific record: https://arxiv.org/abs/2504.15304v2. doi: 10.48550/arXiv.2504.15304

[6] J. Skalse, N.H.R. Howe, D. Krasheninnikov and D. Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, vol. 35, pages 9460–9471, Neural Information Processing Systems Foundation, 2022. doi: 10.52202/068431-0687

[7] S. Casper, X. Davies, C. Shi, T.K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, M. Slocum, U. Anwar, … D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. https://openreview.net/forum?id=bx24KpJ4Eb

[8] V. Conitzer, R. Freedman, J. Heitzig, W.H. Holliday, B.M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf, E. Tewolde and W.S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, pages 9346–9360, PMLR, 2024. https://proceedings.mlr.press/v235/conitzer24a.html

[9] T. Zhi-Xuan, M. Carroll, M. Franklin and H. Ashton. Beyond preferences in AI alignment. *Philosophical Studies*, 182:1813–1863, 2025. doi: 10.1007/s11098-024-02249-w

[10] I. Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020. doi: 10.1007/s11023-020-09539-2

[11] I. Gabriel and G. Keeling. A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*, 182:1951–1973, 2025. doi: 10.1007/s11098-025-02300-4

[12] A. Ecoffet and J. Lehman. Reinforcement learning under moral uncertainty. In *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pages 2926–2936, PMLR, 2021. https://proceedings.mlr.press/v139/ecoffet21a.html

[13] R. Max. Guidelines without guidance: Underdetermination in AI ethics. *AI and Ethics*, 6:Article 323, 2026. doi: 10.1007/s43681-026-01188-y

[14] P.D. Hershock. *Buddhism and intelligent technology: Toward a more humane future*. Bloomsbury Academic, 2021. doi: 10.5040/9781350182301

[15] R.E. Laukkonen, F. Inglis, S. Chandaria, L. Sandved-Smith, E. Lopez-Sola, J. Hohwy, J. Gold and A. Elwood. Contemplative artificial intelligence (arXiv:2504.15125v3) [Preprint, version 3]. *arXiv*, 2025. Version-specific record: https://arxiv.org/abs/2504.15125v3. doi: 10.48550/arXiv.2504.15125

[16] S. Lazar. Governing the algorithmic city. *Philosophy & Public Affairs*, 53(2):102–168, 2025. doi: 10.1111/papa.12279

[17] C.A.F. Rhys Davids (Trans.). *A Buddhist manual of psychological ethics. Being a translation of the Dhammasaṅgaṇī, compendium of states or phenomena* (3rd ed., 1974, reprinted 2004). Pali Text Society.

[18] Y. Karunadasa. *The Theravāda Abhidhamma: Inquiry into the nature of conditioned reality*. Wisdom Publications, 2019.

[19] Nyanaponika Thera. *Abhidhamma studies: Researches in Buddhist psychology* (3rd ed.). Buddhist Publication Society, 1976.

[20] G. Lind. *How to teach morality: Promoting deliberation and discussion, reducing violence and deceit*. Logos Verlag, 2016.

[21] E. Nowak. What is moral competence and why promote it? *Ethics in Progress*, 7(1):322–333, 2016. doi: 10.14746/eip.2016.1.18

[22] F. Schauer. *Playing by the rules: A philosophical examination of rule-based decision-making in law and in life*. Oxford University Press, 1991.

[23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P.F. Christiano, J. Leike and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, vol. 35, pages 27730–27744, 2022. doi: 10.52202/068431-2011

[24] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C.C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, ... T. Scialom. Llama 2: Open foundation and fine-tuned chat models (arXiv:2307.09288v2) [Preprint, version 2]. *arXiv*, 2023. https://arxiv.org/abs/2307.09288v2. doi: 10.48550/arXiv.2307.09288

[25] OpenAI. Expanding on what we missed with sycophancy. OpenAI Research Post-Mortem, May 2, 2025. https://openai.com/index/expanding-on-sycophancy/

[26] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S.R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S.R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang and E. Perez. Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html

[27] A. Siththaranjan, C. Laidlaw and D. Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/7423902b5534e2b267438c85444a54b1-Abstract-Conference.html

[28] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M.D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, ... D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. https://jmlr.org/papers/v23/20-1335.html

[29] J. Eisenstein, C. Nagpal, A. Agarwal, A. Beirami, A. D'Amour, D. Dvijotham, A. Fisch, K. Heller, S. Pfohl, D. Ramachandran, P. Shaw and J. Berant. Helping or herding? Reward model ensembles mitigate but do not eliminate reward hacking. In *Proceedings of the First Conference on Language Modeling (COLM 2024)*, 2024. https://openreview.net/forum?id=5u1GpUkKtG

[30] C. Allen, I. Smit and W. Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3):149–155, 2005. doi: 10.1007/s10676-006-0004-4

[31] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen and A. Bernstein. Implementations in machine ethics: A survey. *ACM Computing Surveys*, 53(6), Art. 132:1–38, 2020. doi: 10.1145/3419633

[32] H.S. Richardson. Specifying norms as a way to resolve concrete ethical problems. *Philosophy & Public Affairs*, 19(4):279–310, 1990. https://www.jstor.org/stable/2265316

[33] T.L. Beauchamp and J.F. Childress. *Principles of biomedical ethics* (8th ed.). Oxford University Press, 2019.

[34] I. van de Poel. Translating values into design requirements. In *Philosophy and Engineering: Reflections on Practice, Principles and Process*, pages 253–266. Springer, 2013. doi: 10.1007/978-94-007-7762-0_20

[35] I. van de Poel. Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, 30(3):385–409, 2020. doi: 10.1007/s11023-020-09537-4

[36] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, pages 220–229, 2019. doi: 10.1145/3287560.3287596

[37] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H. Daumé III and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723

[38] T.K. Gilbert, N. Lambert, S. Dean, T. Zick, A.J. Snoswell and S. Mehta. Reward reports for reinforcement learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2023)*, pages 84–130, 2023. doi: 10.1145/3600211.3604698

[39] J. Mökander, J. Morley, M. Taddeo and L. Floridi. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 27, Art. 44, 2021. doi: 10.1007/s11948-021-00319-4

[40] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv* preprint arXiv:2212.08073, 2022. https://arxiv.org/abs/2212.08073. doi: 10.48550/arXiv.2212.08073

[41] H.L.A. Hart. *The concept of law* (3rd ed.). Oxford University Press, 2012. doi: 10.1093/he/9780199644704.001.0001

[42] Meta. Llama 2 model card. GitHub repository, 2024. https://github.com/meta-llama/llama-models/blob/main/models/llama2/MODEL_CARD.md