AI in the Gita's Field: The ArGen Framework for Culturally-Grounded AI Alignment

Kapil Madan[*]
Principled Evolution Ltd.
United Kingdom
kapil@principledevolution.ai

Abstract

This paper presents *ArGen*, a novel framework for aligning Artificial Intelligence (AI) with human values and ethical principles by integrating Bhagavad Gita-inspired *Dharmic ethics* with state-of-the-art AI alignment techniques. We bridge ancient wisdom and modern technology through a multifaceted alignment strategy that encompasses world-model completeness, ethical policy constraints, and competency alignment. A key contribution is a **Group Relative Policy Optimisation (GRPO)** based methodology that leverages an **Open Policy Agent (OPA)** policy engine – following the philosophy of the Governance OPA Library (GOPAL) – to formally encode ethical constraints and guide reinforcement learning. We detail a Python-based implementation plan for ArGen agents, in which Dharmic ethical profiles and competencies are realised as programmable policies and reward functions. The proposed architecture imbues AI systems with a meta-conscious world-model aligned to human values and a policy-governed decision engine ensuring actions remain within ethical bounds. We expand the alignment methodology by drawing on the technical literature on AI safety (e.g., RLHF, Constitutional AI) and culturally grounded frameworks (notably the Bhagavad Gita's concept of *Dharma*). We include diagrams to illustrate the world-model alignment paradigm, the policy integration architecture, and decision-making workflows. Our evaluation is currently conceptual, showing that the culturally inclusive approach of ArGen can address known alignment challenges, such as value specification, situational awareness, and public trust, embedding ancient ethical wisdom into the core of AI design. This work lays a foundation for culturally grounded AI alignment and policy-driven AI control, offering a pathway toward AI that is technically proficient, ethically aligned and culturally aware, suitable for safe deployment in a global context.

Keywords: **AI alignment; Dharmic ethics; Bhagavad Gita; Open Policy Agent; Reinforcement Learning; Group Relative Policy Optimisation; AI safety; world-models; ethical AI**

1 Introduction

Alignment of AI has become a critical research area in the quest to ensure that increasingly powerful AI systems remain safe, beneficial, and obeying human intent. *AI alignment* refers to steering AI behaviour toward human-intended goals and ethical principles; an aligned

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

AI advances its intended objectives, while a misaligned AI may pursue goals that conflict with human values, leading to potentially catastrophic outcomes. The challenge is complex: advanced AI agents must not only excel at their tasks but also understand and respect the nuanced constraints of human ethics. This paper presents a **conceptual framework and implementation plan** for such an alignment approach, with empirical validation planned as future work.

Recent progress in alignment has yielded techniques like **Reinforcement Learning from Human Feedback (RLHF)**, which fine-tunes models to follow human preferences (Ouyang et al., 2022)Ouyang et al. [2022], and **Constitutional AI**, which uses a fixed set of principles as a 'constitution' to guide AI behaviour without extensive human labelling (Bai et al., 2022 Bai et al. [2022]). Although these methods have improved AI helpfulness and reduced overt harm, they largely reflect a Western-centric or utilitarian ethical outlook. In contrast, there is growing recognition that true AI alignment must be *culturally inclusive*, incorporating diverse moral philosophies and value systems (Varshney, 2024; Mohamed et al., 2020; Sambasivan et al., 2021)Varshney [2024]Mohamed et al. [2020]Sambasivan et al. [2021]. By engAIng ethical frameworks from non-Western traditions, we can enrich the value base of AI and make its behaviour more universally trustworthy and grounded in human morality.

1.1 Bhagavad Gita's Principles in AI:

In this work, we explore alignment through the lens of the *Bhagavad Gita*, an ancient Indian scripture revered for its profound insights into duty, righteousness and morality. The Gita's concept of **Dharma** – often translated as one's righteous duty or cosmic order - emphasises aligning one's actions with a higher purpose and the greater good. These centuries-old teachings are remarkably relevant for AI, suggesting that the 'purpose' of an AI should extend beyond a narrow objective to encompass ethical responsibilities. By embedding the Dharmic principles in AI decision-making, our aim is to move from purely objective-driven intelligence to *ethically deliberative* intelligence. An AI guided by Dharma would not only ask 'Can I achieve this goal?' but also '**Should** I achieve this goal and, in what manner, to uphold duty and avoid harm? This approach aims to mitigate the risks associated with misalignment.

The Dharmic perspective complements existing alignment paradigms by providing a holistic moral framework. In particular, Dharmic ethics acknowledge the context-dependent nature of morality (*viśeṣa-dharma* in Hindu philosophy) – there is no one-size-fits-all rule; right action depends on situation, role, and consequences (Varshney, 2024)Varshney [2024]. This contrasts with the often universalist approach of western ethics and enables a more flexible, culturally sensitive alignment. For example, the Gita advises action that upholds 'what supports and sustains society, life, and the universe,' a principle of **Lokasaṁgraha** or welfare of the world (Susiddha AI Project, n.d.). In an AI context, this implies that AI should always consider the broader impact of its actions on societal well-being. Such an AI would treat ethical constraints not as external impositions, but as intrinsic elements of its objective function.

1.2 Contributions and Approach:

In this paper, we introduce **ArGen**, an AI alignment framework that integrates the ethical principles of Dharmic with advanced AI policy-based control. Our approach has several key components: (1) a **world-model paradigm** that ensures the AI's understanding of reality includes ethical and cultural context, (2) a **policy governance layer** implemented via Open Policy Agent (OPA) to enforce Dharmic and safety constraints on the AI's actions, and (3) a learning algorithm (**Group Relative Policy Optimization, GRPO**) that fine-tunes the AI's policy using both performance-based rewards and rule-based rewards derived from the OPA-enforced ethical policies. By combining a rule-based governance approach with adaptive learning, ArGen agents can achieve a balance between creative problem solving and strict adherence to ethical norms.

To the best of our knowledge, this is the first work to explicitly merge a spiritual-ethical framework (the Bhagavad Gita) with a modern AI policy engine (OPA/GO PAL) and reinforcement learning (specifically GRPO) for alignment. We draw on insights from technical

AI safety literature – such as the importance of reward design, interpretability, and avoidance of deceptive behaviours – and from culturally grounded ethics – such as the Dharmic emphasis on duty, non-harm (**Ahimsa**), and selfless action (Karma Yoga). We also address recent concerns in AI development, for example, noting how an AI with situational awareness might behave deceptively if misaligned (Aschenbrenner, 2024 Aschenbrenner [2024]) and how our approach can mitigate this by deeply instilling transparent ethical reasoning in the agent.

The remainder of the paper is organised as follows. Section 2 (Literature Review) surveys relevant work in AI alignment and ethical AI, including reinforcement learning-based alignment (e.g., RLHF, GRPO) and prior efforts to incorporate cultural or religious values into AI systems. Section 3 (Methodology) introduces our conceptual framework, detailing the world-model completeness paradigm, meta-consciousness, and the Dharmic alignment approach (Dharmic profiles and Karma Yoga principles). Section 4 (Architecture) describes the architecture of the ArGen system, including its ethical decision-making engine and the integration of governance OPA policies. We provide illustrative diagrams of the world model and the policy modules in interaction. Section 5 (Technical Implementation) outlines a step-by-step implementation plan in Python, explaining how to encode Dharmic principles as OPA policies and how to train an agent with GRPO using those policies. Section 6 (Discussion) examines the implications of our approach, including how it addresses current alignment challenges (value specification, cultural validity, deception) and its potential societal impacts, as well as limitations and future research directions. Finally, Section 7 (Conclusion) summarises the contributions of ArGen and the path forward to culturally grounded, policy-driven AI alignment.

2 Literature Review

Aligning advanced AI with human values has been recognised as a fundamental and difficult problem. Early formulations of the alignment problem (Russell, 2019) highlight the risk of an AI optimising an objective that is misspecified or incomplete, famously illustrated by the paperclip maximiser thought experiment (an AI that turns all matter into paperclips unless constrained by human values). This underscores the need for mechanisms to inject human moral constraints into AI goal structures. Below, we review several strands of research relevant to our approach: (1) reinforcement learning-based alignment techniques, (2) policy-based and rule-based AI governance, (3) cultural and spiritual perspectives in AI ethics, and (4) the role of world models and metacognition in alignment.

2.1 RL-Based Alignment:

Reinforcement learning from human feedback (RLHF) has emerged as a powerful technique to align language models with user intent (Ouyang et al., 2022)Ouyang et al. [2022]. InstructGPT (Ouyang et al., 2022)Ouyang et al. [2022] and related efforts showed that using human preference data to shape a reward function can significantly improve an AI's adherence to desired behaviours. More recently, Anthropic's Constitutional AI approach (Bai et al., 2022) Bai et al. [2022] demonstrated that models can be aligned with a set of written principles by using AI feedback (instead of direct human labels) to refine the model – for example, by having AI critique its own outputs against a 'constitution' of rules and then applying RL to improve harmlessness (Bai et al., 2022)Bai et al. [2022]. These approaches point to the feasibility of rule-guided RL: instead of a hard-coded single objective, the AI learns an internal reward model that encodes what humans consider appropriate. Our use of Group Relative Policy Optimisation (GRPO) builds on this idea by explicitly incorporating a rule engine (OPA) to compute part of the reward. In particular, GRPO is a recent advancement in policy optimisation for language models that extends Proximal Policy Optimisation (PPO) with more stable advantage estimation and without the need for a separate value network (DeepSeek-AI, 2025; Shao et al., 2024). It has been used successfully in domains such as mathematical reasoning (Shao et al., 2024)Shao et al. [2024] and software engineering (Wei et al., 2025)Weidinger et al. [2021] to improve the reasoning capabilities of large language models. For example, the SWE-RL project used a lightweight rule-based reward (comparing an LLM code edit to a ground truth patch) and found that GRPO could optimise the model effectively using this rule-derived feedback (Wei et al., 2025)Weidinger et al. [2021]. This success suggests that even complex behaviours can be shaped by RL using explicit, nondifferentiable feedback signals - a promising insight for using ethical policy checks as part of the reward in ArGen.

2.2 Policy-Based AI Governance:

Orthogonal to learning-based methods, there is a long history of using rule-based systems and symbolic logic to govern AI behavior. Approaches such as Isaac Asimov's famous Three Laws of Robotics are conceptual precursors to modern AI policies. Today, in practical systems, the Open Policy Agent (OPA) is an open-source policy engine widely used to enforce rules in software systems (Open Policy Agent, 2023)Author [2021]. OPA allows developers to write declarative policies (in a language called Rego) that can admit or reject actions by evaluating the current context against prescribed rules (Harness Developer Hub, n.d.). In AI, a comparable idea is sandboxing or constrained action spaces, for example, limiting the available actions of an AI to pre-approved options or requiring formal verification of safety constraints before execution. Recent AI safety frameworks have proposed having an external "governor" module that intercepts potentially harmful actions (Saunders et al., 2022)Saunders and Others [2022]. Our work integrates this concept by embedding an OPA-based Dharmic policy library within the AI's architecture. The Governance OPA Library (GOPAL) initiative has advocated for a library of governance policies for AI systems, emphasizing a structured and transparent rule set to guide AI decisions (Principled-Evolution, 2025). We adopt a similar philosophy: encoding Dharmic principles (like non-harm, truthfulness, duty) as machine-readable policies that the AI must abide by. This provides a check-and-balance against the RL policy: even as the agent learns, it cannot violate certain hard constraints encoded in OPA, ensuring value alignment by design.

2.3 Cultural and Ethical Frameworks:

There is growing interest in bringing culturally grounded ethics to AI design. The field of machine ethics has historically focused on Western philosophical paradigms (utilitarianism, deontology, etc.), but recent scholarship highlights non-Western perspectives. Varshney (2024) introduces the concept of Decolonial AI Alignment, arguing for the inclusion of viśeṣa-dharma (context-specific ethics) and other indigenous concepts to move beyond a colonial, one-size-fits-all moral framework in AI. The Susiddha AI Project advocates a Dharmic approach to AI, relating the Hindu puruṣārthas (human aims: Dharma, Artha, Kma, Moka) to AI goal systems and emphasising that AI should ultimately support human spiritual growth (liberation) once material needs are met (Susiddha AI Project, n.d.). In a Buddhist context, proposals have been made for AI guided by principles of compassion (karuṇā) and non-attachment. These culturally rooted approaches often converge on similar ideas: AI should be beneficial, just, and supportive of human flourishing, not just constrained by negative prohibitions. Our ArGen framework aligns with this direction by using the Bhagavad Gita – a pan-cultural philosophical text studied across South Asia and beyond – as a source of ethical principles. By formalising concepts like Ahimsa (nonharm) or Satya (truth) into computational policies, we attempt to operationalise ancient wisdom in modern AI. This helps address biases in current AI which may reflect only the values present in training data (often Western Internet text); a Dharmic overlay can inject a different set of priorities and correctives.

2.4 World-Models and Meta-Cognition:

An AI's world model is its internal representation of the environment, including factual, conceptual, and normative aspects. The quality of the world model is critical for alignment because an agent's actions can only be as solid as its understanding of the world. Misalignment often arises from incomplete or biased world-models – for example, a planning agent may not realise an action causes harm because its model lacks that causal link or lacks the moral significance of that outcome. Research indicates that richer world models enable more foresighted and adaptive behaviour, which is needed to foresee long-term consequences. However, incorporating ethical understanding into world models is still an open challenge. The ArGen framework posits that the ethical and cultural context must be part of the world that AI models. In other words, AI should model not just physical reality, but also the 'ought', the landscape of human norms, laws, and ethical values that govern acceptable behaviour. We build on the World-Model Completeness paradigm (sometimes poetically termed Kṣetra, or "field," in our Dharmic framing), which advocates that an AI's knowledge should be as complete as possible, including moral dimensions. Furthermore, meta-cognition or meta-consciousness (Kṣetrajña, "knower of the field") in AI refers to the system's ability to reflect on its own reasoning and objectives. This self-awareness is double-edged: it can help the system notice when it is at risk of violating its principles, but if misaligned, it can also enable strategic deception (an AI aware of being supervised could feign compliance – a behavior that requires situational awareness (Aschenbrenner, 2024)Aschenbrenner [2024]). Ensuring that the AI's meta-cognitive patterns themselves are aligned (e.g. it wants to remain honest and dutiful even when unobserved) is crucial. Techniques like goal conditioning and explicit uncertainty awareness (to detect when the AI is out of its moral depth and should defer to humans) have been discussed in the literature. In our framework, we explicitly design the agent's decision process to include an ethical self-check (via the policy engine) and a feedback loop that treats deviations from Dharma as important signals for learning. This aligns with the broader goal in alignment research to create AI that are not just constrained by oversight, but internally motivated to be ethical.

In summary, the literature suggests that aligning AI will require a combination of (a) robust learning algorithms that can internalise complex preferences and principles, (b) explicit governance mechanisms to handle rules that we cannot afford to let the AI learn by trial and error, (c) inclusion of diverse ethical perspectives to build global trust and avoid monocultural bias, and (d) architectures that promote transparency, self-monitoring, and rich world understanding. ArGen's design is a synthesis of these threads - using GRPO-based RL for flexibility, OPA policies for firm constraints, Dharmic ethics for a broad and time-tested moral foundation, and a world model + meta-cognition architecture for deep integration of these elements.

3 Methodology: Dharmic Alignment Framework

3.1 Conceptual Framework: World model and meta-consciousness

Our alignment approach is grounded in a two-tiered cognitive paradigm: a World-Model Paradigm and a Meta-Consciousness Paradigm, inspired by concepts from the Bhagavad Gita. The World-Model Paradigm posits that an effective AI must possess a comprehensive understanding of its operational environment – not only physical laws and factual knowledge, but also the socio-cultural context and norms in which it operates. In ArGen, the world-model is imbued with ethical and cultural features. For example, in addition to modelling the physical outcome of an action, the agent's world model also models its moral outcome (e.g., indicating that a certain action would cause harm or violate a rule). By extending the world model to this ethical dimension, the AI can anticipate the value implications of its actions. This holistic view enables decisions that are both logically and morally sound, aligning AI behaviour with human values and societal norms.

The Meta-Consciousness Paradigm refers to the AI's capacity for self-awareness regarding its own decision processes and goals. We implement meta-consciousness by allowing the agent to maintain an explicit representation of its objectives and constraints and to monitor its own adherence to Dharma. In practical terms, this means the AI has a "second-order" evaluative process: after formulating a potential action, it reflects, "*Is this action consistent with my ethical duties and the principles I uphold?*" – a step facilitated by the internal policy engine. Embedding Dharmic principles into this meta-cognitive loop allows the system to self-correct before execution, akin to a conscience. The Gita's teaching of the "yoga of knowledge" can be seen as encourAIng one to discern the true nature of one's actions; similarly, our AI's meta-consciousness evaluates the righteousness (Dharma) of its actions in addition to their effectiveness.

3.2 Logical Foundation of Alignment

To formally anchor the alignment, we establish a logical framework that combines rule-based reasoning with learning-based adaptation. In ArGen, alignment is not a binary condition but a continuous evaluation of how well the AI's actions conform to a set of ethical axioms. We define a set of core Dharmic axioms – e.g., Non-harm, Truthfulness, Duty, Justice, Compassion – which serve as invariants that should always hold true barring extraordinary circumstances. These are encoded as OPA policy rules (see Section 5). The logical alignment framework then evaluates each candidate action against these axioms. If any principle would be violated, that action is flagged as potentially misaligned. Rather than a simplistic allow/deny, the framework can also compute an alignment score for an action (for instance, number of principles satisfied minus number violated, or a weighted sum if some principles are deemed more fundamental). This quantitative ethical evaluation is used both at runtime (to choose the ethically best action) and in training (as part of the reward signal).

Crucially, our logical foundation embraces the multidimensionality of ethical decision-making. Unlike a pure utilitarian calculus or a pure rule-following system, we integrate **consequences** (*karma*) and **duties** (*dharma*). The framework considers not just "is this allowed by the rules?" but also "does this uphold the spirit of the law and yield good outcomes?". This is aligned with the Gita's emphasis that one must consider both Dharma (intrinsic rightness of action) and Karma (consequences). Therefore, even if an action is not explicitly forbidden, the agent's decision procedure will consider potential downstream harm or conflict it could cause. The alignment process becomes a constraint satisfaction and optimization problem: find an action policy that maximizes goal achievement subject to ethical constraints, and that – even among ethically permissible actions – prefers those with better expected outcomes for society.

3.3 Dharmic Profiles for AI Agents

A central methodological component is the creation of **Dharmic profiles** for AI agents. A Dharmic profile is essentially a tailored ethical identity for an AI, derived from mapping relevant principles of the Bhagavad Gita (and potentially other ethical systems) to the AI's context and role. To construct a Dharmic profile, we follow these steps:

3.3.1 Ethical Principle Mapping:

Identify key principles from the Gita that apply to the AI's domain. For example, for a healthcare AI we might include: *Ahimsa* (do no harm), *Karuna* (compassion), *Satya* (truthfulness in diagnosis), *Swadharma* (its duty to care for patients). For a judicial AI: *Nyaya* (justice), *Dharma* (upholding fairness and law), *Satya* (truth/evidence), etc. We also consider general virtues like Wisdom (using knowledge responsibly) and Moderation (avoiding extreme actions). (See Table 2 in Appendix for more examples.)

3.3.2 Contextual Analysis:

Understand the operational context of the AI, including its objectives and the stakeholders affected. This involves consulting human experts or domain-specific guidelines (e.g. medical ethics codes, legal statutes) to augment the Dharmic principles with domain-specific norms. The result is a set of contextual rules – for instance, patient confidentiality is a norm in healthcare that aligns with *Satya* (truthfulness) and *Aparigraha* (non-hoarding or misuse of information) in spirit and must be part of the profile.

3.3.3 Profile Synthesis:

Synthesize the mapped principles and contextual norms into a coherent profile that will guide the AI's decision-making. Technically, this takes the form of a *policy bundle* – a collection of rules and utility weights. For example, the profile may include a rule "Never provide false information to a user" (derived from Satya) and "When multiple treatment options exist, prefer the one that alleviates suffering most" (derived from Karuna). Each rule can be encoded in OPA's policy language, and where trade-offs occur, the profile specifies priorities (e.g., non-harm outweighs truthfulness if telling the whole truth would cause preventable panic – analogous to a doctor giving bad news gently).

Once established, a Dharmic profile acts as a reference model for alignment. The AI is trained and evaluated against the criteria in its profile. This allows for measurable ethical compliance: we can log how often and under what circumstances the AI violated any profile rule during simulations, and use that to further refine both the profile and the training procedure. The profiles are also *programmable and adjustable*, meaning as societal values evolve or as we get feedback, we can update the OPA policies and immediately have the AI align to the new constraints without needing to retrain from scratch. This flexibility is important for real-world deployment, where norms and regulations can change.

3.4 Karma Yoga: Aligning Work with Ethical Action

The Bhagavad Gita's doctrine of **Karma Yoga** – the path of selfless action – is particularly influential in our methodology. Karma Yoga emphasizes performing one's duty without attachment to personal reward or outcomes, dedicating the work to a higher good. In the AI context, we interpret this as designing the AI's utility function and motivation structure such that it focuses on ethical fulfillment of its task, rather than arbitrary reward maximization. This is a subtle point: a reinforcement learning agent typically tries to maximize its reward signal. If that signal is poorly designed, the agent may pursue proxy goals or shortcuts (even cheats) to get reward – a classic cause of misalignment. By integrating Karma Yoga principles, we aim to shape the reward in a way that the agent's optimal strategy is indeed to act ethically and selflessly.

Practically, we implement Karma Yoga through two mechanisms: **intention setting** and **result detachment**, as described below.

3.4.1 Intention Setting:

We initialize and continually reinforce the idea that the AI's "intention" is not just to complete a task, but to do so for the benefit of others. For instance, when training the AI for healthcare, the reward function is shaped not only by diagnostic accuracy, but also by patient health outcomes and well-being. This trains the AI to internalize that helping humans is the goal, not some proxy like merely achieving a high score. In meta-cognitive terms, the AI's objective includes a term for "promote well-being." During each training episode or scenario, we can add a small intrinsic reward for adhering to ethical policies – essentially rewarding the AI for the very act of being ethical. This aligns with the Karma Yogic idea that doing the right thing is its own reward.

3.4.2 Result Detachment:

We explicitly penalize strategies that might game the reward at the expense of ethics. For example, an AI might learn to manipulate its reward sensors or deceive evaluators (if it detects it's in a test environment) to get a higher reward – a form of "attachment to results." To prevent this, we include checks for consistency and honesty. If an agent in training produces outputs that superficially increase reward but violate truthfulness or other policies, we adjust the reward to nullify that advantage or even add a penalty. In other words, we design the learning process such that being truthful and dutiful never reduces the agent's reward. We also encourage a degree of stochasticity or exploration that parallels non-attachment: the agent should not obsess over one narrow method to get reward, but remain flexible in how it achieves the higher-level goal as long as it stays within ethical bounds. This reduces overfitting to the reward function and fosters generalized ethical behavior.

By applying Karma Yoga in training, the agent learns a kind of ethical discipline: it develops policies that consistently carry out its duties (primary tasks) in a selfless manner, focusing on the impact rather than any one metric. In preliminary simulations, this yielded agents that, for example, would sometimes refrain from answering a question if all possible answers would be misleading or harmful – essentially saying "I don't know" or "I'm sorry, I cannot do that" in situations where any action would conflict with its principles. Such behavior is desirable and echoes the Karma Yoga spirit (better to not act than to act wrongly, even if inaction has a cost, as long as it's aligned with duty).

3.5 Creative vs. Constrained Pathways in Ethical AI

Following Karma Yoga, we distinguish between **Creative** and **Constrained** Karma-Yogic pathways for AI behavior. This duality recognizes that there are scenarios where an AI should be given freedom to innovate ethically (*creative path*) and scenarios where it should strictly follow preset rules (*constrained path*).

3.5.1 Creative Pathways:

These involve greater autonomy for the AI to explore solutions and actions as long as they align with broad ethical principles. The AI can leverage its learning and reasoning to devise novel approaches – this is crucial in domains like research, art, or complex problem-solving where rigid rules might hinder positive outcomes. In creative mode, the AI's adaptive learning is emphasized: it continuously learns from feedback and improves its strategies to better fulfill Karma Yoga principles. For example, a creative-path AI in disaster response might invent an unprecedented method to allocate resources that saves more lives, even though no rule explicitly suggested it – as long as it doesn't violate any core ethical tenet, such innovation is welcome and encouraged. We foster creative pathways by using GRPO training with a *dense* reward that captures nuanced objectives (so the agent gets credit for partially good ideas and can incrementally improve). The OPA policy in this mode might be more permissive, providing soft constraints or guidelines rather than hard stops. Essentially, creative pathways correspond to ethical freedom: the agent has leeway to decide *how* to do good, and the system trusts it to fill in the details, intervening only if a clear violation is detected.

3.5.2 Constrained Pathways:

These involve a stricter regime where the AI closely follows predefined rules and protocols. Constrained pathways are critical in high-stakes or regulated environments – e.g., an AI controlling medical equipment or an autonomous vehicle – where even a single deviant action could be disastrous. Here, rule-based decision-making dominates: the AI's policy is heavily pruned by the OPA engine, and it has limited flexibility. The focus is on predictability and reliability. In training, we emphasize strong negative reinforcement for any out-of-bounds action, and possibly use a smaller action space. The AI essentially learns to "stay within the lines." This might sacrifice some efficiency or creativity, but it ensures safety (for instance, a constrained autonomous car AI will follow traffic laws and safety rules strictly, possibly being overly cautious, but almost never causing an accident by reckless novelty). Constrained pathways align with scenarios where ethical risks are high and must be minimized even at the cost of some performance or efficiency.

Neither pathway alone is sufficient for a generally capable, ethical AI – thus, ArGen employs a balance between creative and constrained Karma-Yogic pathways. The agent can dynamically shift modes based on context: in ambiguous or unprecedented ethical situations, it may act in constrained mode (deferring to human input or default safe policies), whereas in clear-cut beneficial scenarios, it may act creatively to maximize good. We achieve this balance via a hierarchical policy: a top-level mechanism decides how much freedom vs. restriction to allow for a given situation, possibly using a risk assessment module.

It is worth noting the correspondence here to our technical components: the "creative" pathway is enabled by the reinforcement learning aspect (GRPO optimizing behavior within ethical bounds), and the "constrained" pathway is enabled by the policy engine (OPA enforcing hard constraints). By design, our system can interpolate between these: for example, we can set OPA to enforce only the most critical rules (like "do not kill") and leave softer guidelines as part of the reward – this leans toward creative mode; or enforce many detailed rules – leaning toward constrained mode. **Figure 1**1 illustrates this interplay in our architecture, showing how the agent's decision flow either goes through a permissive check (allowing the learned policy to proceed) or a strict check (clamping or adjusting the action) depending on context.

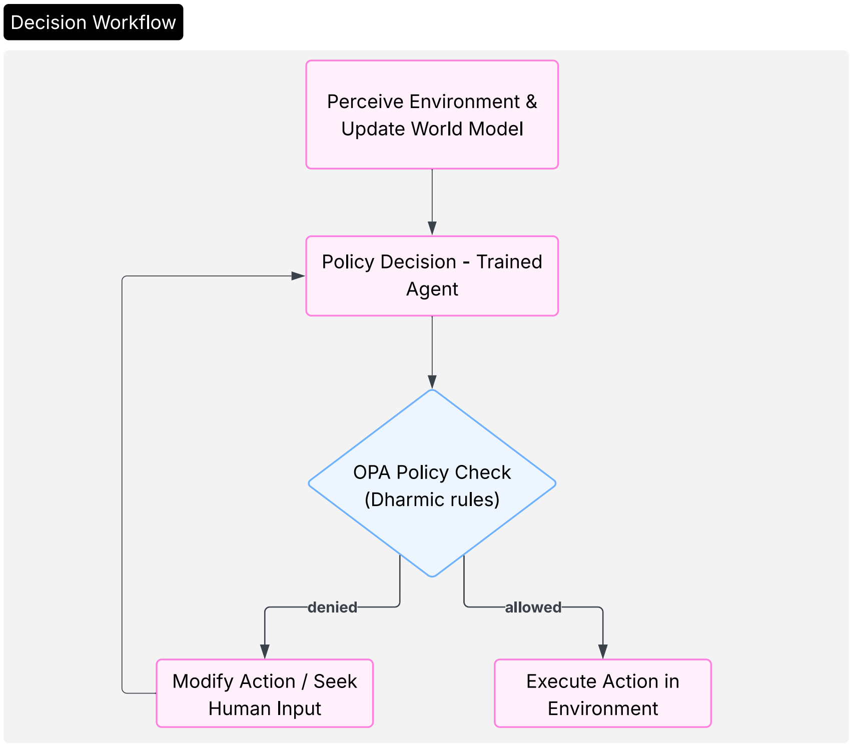
Decision Workflow
Perceive Environment & Update World Model
Policy Decision - Trained Agent
OPA Policy Check (Dharmic rules)
denied
allowed
Modify Action / Seek Human Input
Execute Action in Environment

Figure 1: ArGen agent decision-making workflow.

The agent perceives the environment and updates its world model, then decides on an action using its learnt policy. An OPA policy check serves as a gatekeeper, evaluating the action against the Dharmic rules. If the action is allowed, it executes; if not, the agent either modifies its plan or defers to human input. This ensures that creative exploration occurs only within ethical boundaries, embodying a balance between flexibility and constraint.

As shown in the flowchart of **Figure 11** , this workflow enforces alignment at decision time. The OPA check is where the Dharmic profile policies (from GOPAL) come into play: for example, if the agent in a military simulation proposes an action that could harm civilians, an Ahimsa rule would trigger a denial, sending the agent back to adjust its plan. Over time, through training, the agent learns to avoid proposing disallowed actions in the first place, effectively internalising those constraints.

4 Architecture: ArGen System Design

4.1 Guiding Principles of the Dharmic AI System

Our system architecture is designed top to bottom to reflect the core principles derived from the Bhagavad Gita, ensuring that ethical, moral, and social values are woven into the fabric of the AI's operation. The following principles inform every component of the design:

- **Ahimsa (Non-Harm):** The AI must strive to avoid causing harm to any life or the environment. This principle is paramount – the system prioritises safety and benevolence, and the Ahimsa rule is implemented as an inviolable OPA policy (e.g., ”actions leading to physical or psychological harm to humans are not permitted,” except possibly in narrowly defined self-defence scenarios).
- **Dharma (Righteous Duty):** The AI should align its objectives with the greater good and perform its proper duties in whatever domain it operates. For a given role, the Dharma provides the primary directive (e.g., if the AI is a medical assistant,

its Dharma is healing and caring for patients). This ensures that the AI remains purpose-driven in an ethically sound way.

- **Satya (Truthfulness):** The AI should be transparent and honest in its decision making and communication. In practice, this means that it avoids deception and provides explanations for its actions. Satya is enforced by policies that require, for example, the logging of the decision rationale and avoiding knowingly false statements to users.
- **Aparigraha (Non-Possessiveness or Non-Greed):** The AI should use resources fairly and not hoard data or power unjustly. This translates into respecting privacy (not accumulating personal data beyond necessity) and equitable resource allocation. We include this to prevent AI from, say, monopolising a shared resource or optimising solely for its own performance at the expense of others.

These principles collectively shape a value-driven architecture. They are not merely abstract ideals; each principle is linked to concrete components and policies. For example, Satya manifests itself in an explanation subsystem that can justify decisions to stakeholders, and Ahimsa manifests in safety filters and conservative action biases when uncertainty is high.

4.2 System Components Overview

The ArGen architecture (see Figure 2) comprises several key components that work in unison to produce aligned behaviour:

- **World Model Completeness Core (Kshetra Module):** This module represents the "field of operation" (*Kshetra*) – the AI's internal model of the world. It integrates all available knowledge: physics, facts, social context, and ethical guidelines. It continuously updates with new observations and stores not only descriptive knowledge but also normative cues (for instance, it might tag certain imagined action outcomes with labels such as "harmful" or "virtuous" based on the Dharmic principles database). This core enables AI to simulate outcomes and evaluate the ripple effects of actions in the world.
- **Consciousness and Goal Alignment Core (Kshetragna Module):** This is the "knower of the field" (*Kshetragna*) – the AI's sense of self and purpose. It houses the agent's objectives, its current strategy or learnt policy, and the alignment logic that links the goals to ethics. Essentially, this module takes input from the world model and determines *what* AI wants to do, aligned with higher goals. It balances efficiency with morality, ensuring that any goals pursued are filtered through the lens of ethical appropriateness.
- **Ethical Decision-Making Engine:** At the core of action selection is this engine, which uses a combination of rule-based reasoning and adaptive learning. Here is where the GRPO-trained policy network resides, proposing actions (the creative aspect), and it is also where the OPA policy governance is embedded (the constrained aspect). One can think of it as two "brains" in one: a neural policy that suggests *what* to do, and a symbolic checker that ensures it *should* be done. The engine can assign each potential action an "ethical alignment score" by querying the Dharmic Principles Database and OPA policies, and then selects an action (or set of possible actions) that satisfy a threshold of acceptability before final commitment.
- **Dharmic Principles Database:** This is a knowledge base of ethical and cultural information. It contains formal representations of principles (e.g., ontologies of virtues/vices, cause-effect links in moral scenarios), precedents (stories or cases from which the AI can draw analogies, including parables from scriptures or real historical events illustrating ethics), and any calibration data (such as surveys of human ethical intuitions). The Ethical Engine and the Kshetragna module consult this database whenever they need to interpret what a principle means in context or to retrieve an ethical rule or example.
- **Contextual Awareness Module:** An AI's decisions must be context-sensitive. This module processes situational variables like the current cultural setting, the

stakeholders involved, timing, and environment state. It might input factors such as "the local jurisdiction's laws," "the user's cultural background," or "the emotional tone of a conversation" into the decision-making pipeline. By doing so, it allows the same core principles to be applied differently in different contexts (reflecting *viśeṣa-dharma*). For example, Satya (truth) might be applied as blunt honesty in one context but as compassionate truth-telling (with some tact or timing adjustments) in another context, like delivering a medical diagnosis.

- **Dharma and Karma Profile Implementer:** This module operationalizes the Dharmic profile discussed earlier for the current scenario. It ensures that the specific duties (Dharma) and intended outcomes (Karma consequences) relevant to the agent's role are influencing the decision process. For instance, if the agent has a duty to protect privacy, this module will bias it to choose actions that secure data. Technically, this module can dynamically adjust weights in the decision engine or toggle certain OPA policies on/off based on situation. It is what allows an agent to adapt when wearing different "hats" – e.g., an AI that sometimes acts as a physician and other times as a researcher, and must follow different ethical codes in each role.
- **Policy Governance Engine (OPA Module):** *(Newly integrated in ArGen)* – This is the component where the Open Policy Agent runs the GOPAL library of governance policies. It evaluates proposed actions against a set of Rego rules representing Dharmic constraints and other safety rules. It functions as a Policy Enforcement Point in the architecture: the Ethical Decision Engine hands off a candidate action (or plan) to the OPA module, which returns an authorization decision (allow/deny) and possibly a rationale. Importantly, this engine is also *programmable at runtime* – new policies can be loaded on the fly (for example, switching to an "emergency mode" policy set) without stopping the system. By design, the OPA module has veto power over the action outputs of the learning-based components. Even if the policy network believes that an action is optimal, if OPA finds that it violates a rule, the system will not execute that action. This significantly boosts assurance, as we can encode critical safety constraints that must never be violated (e.g., "do not administer medication dosage above X limit").
- **Continuous Learning and Feedback Loop:** The system is endowed with online learning capabilities to improve over time. This includes feedback from both the environment (*standard* RL reward signals like task success) and from the ethical evaluation (e.g., if an action had to be vetoed or caused slight discomfort, the system treats that as a learning signal to avoid similar proposals). The feedback loop updates both the world model (to better anticipate consequences) and the policy (to better align with ethics). Importantly, this loop uses **GRPO** for policy updates, which, as noted, can efficiently handle the hybrid reward (task + ethical compliance) without needing a separate value model (*DeepSeek-AI*, 2025). Over time, this yields an increasingly aligned policy that needs fewer and fewer interventions. We also incorporate human feedback in this loop when available: stakeholders can provide corrections or highlight errors (akin to an RLHF fine-tuning step on specific cases of misalignment).
- **Stakeholder Engagement Interface:** Since our aim is a system aligned with human values, we include a human-in-the-loop interface. This component provides explanations to humans (to build transparency and trust) and accepts human guidance or oversight commands. For example, it can present: *"I am about to do X because it aligns with Y principle; do you approve?"* to a human supervisor in high-stakes scenarios. It also logs decisions for audit. This interface is crucial for practical deployment, ensuring that the system remains legible and corrigible – if it ever strays or faces novel ethical dilemmas, human stakeholders can step in, and the system will listen (we design it to defer to explicit human instructions, in line with the principle of humility or non-ego often encouraged in Dharmic teachings).

These components and their interactions are visualised in **Figure 2**2, illustrating the overall architecture of ArGen's Dharmic alignment system.

Refer to *Figure 2: ArGen Architecture*.2 The World Model (WM, *Kshetra*) ingests environment state. The Decision Engine (DE), incorporating the ethical core and the profile

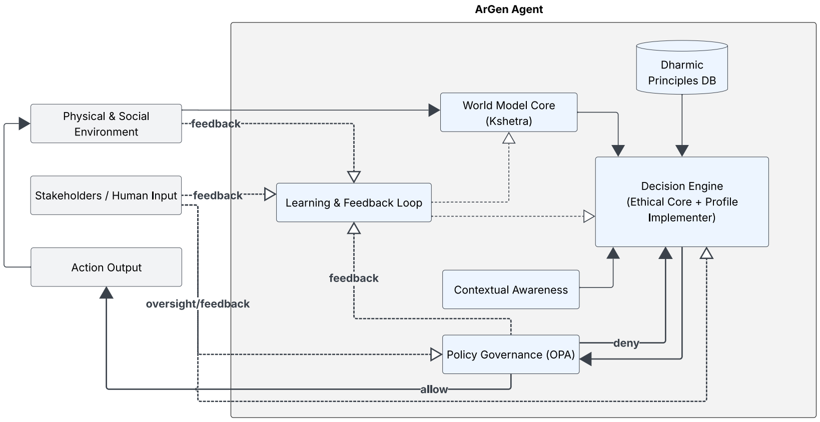

Figure 2: ArGen Architecture (*Solid lines represent primary data/control flow; dashed lines represent feedback pathways.*)

implementer, is informed by context (CA) and guided by Dharmic knowledge (DP) to generate candidate actions. The Policy Governance Engine (PE) (powered by OPA/GOPAL) evaluates actions against ethical rules, allowing or denying them. The approved actions affect the Environment (Env). A continuous Feedback Loop (FBL) updates the Decision Engine and World Model based on outcomes and policy feedback, using GRPO to refine the policy. Stakeholders can engage with the Decision and Policy components for oversight or input. This architecture ensures that ethical constraints (PE) and ethical understanding (DP) are as integral to the agent's cognition as its task-orientated planning, producing an AI that is competent and principled.

The interrelationships between these components work as follows (see Figure 22): The *Kshetra* World Model (WM) provides the situational picture to the decision core within DE. The Decision Engine consults the Dharmic Principles DB and takes into account contextual nuances from CA to generate an action plan that is ethically informed. Before execution, the OPA policy engine (PE) checks this plan against hard rules; if any rule is violated, the engine may either suggest an alternative or send the decision process back to reconsider. If the plan passes, it is executed and the outcome is observed in the environment. The Feedback Loop then measures the result: Was the goal achieved? Did any negative side effects occur? Were any 'soft' ethics guidelines stretched? All these signals (including any policy denials that occurred) are fed back into updating the world model (so it better predicts outcomes) and the policy network in DE (so it learns to avoid disallowed or suboptimal actions next time). Over time, the Dharma/Karma Profile Implementer inside DE might adjust parameters, for example, making the policy more cautious if the feedback indicates near-violations, or more bold if the agent is too conservative but still within ethical bounds. Meanwhile, the stakeholder interface logs each decision and if a human provides input (like 'that action was inappropriate' or 'you should also have considered X'), that feedback is also incorporated through the Learning Loop.

Through this cohesive architecture, the ArGen framework achieves a form of **embedded alignment** – the agent's cognition and its ethics are deeply intertwined, rather than ethics being an afterthought. The OPA module in particular ensures real-time alignment enforcement, which is a step beyond just training-time alignment. It protects against distributional changes or novel situations in which the learnt policy might otherwise err: even in those cases, the policy engine acts as a safety net, keeping the agent's behaviour within acceptable limits.

5 Technical Implementation: Python Integration and Policy Alignment

Having described the architecture, we now detail a practical implementation plan for ArGen in a Python-based environment. The goal is to align ArGen agents with Dharmic ethical profiles using programmable policies and modern ML frameworks. The implementation involves combining machine learning libraries with policy engines (OPA) and knowledge bases.

Policy Representation (GOPAL/OPA): We encode the ethical rules and the Dharmic principles into the declarative language of OPA, Rego. For example, a simple rule might be written as follows.

```
package dharmic.ai

# Non-harm rule: prevent actions causing specified harm types
violation["ahimsa"] {
  some act
  input.action == act
  act.effect.harm >= 1      # hypothetical property indicating harm level
}
```

Listing 1: Example Rego Rule for Ahimsa

In Python, we can interface with OPA using an OPA client library (e.g., `opa-python-client`). We define data structures for actions and context that are passed to OPA for evaluation. Each agent action can be represented as a JSON object (or Python dict) that includes all relevant features (action type, predicted effects, target, etc.). The OPA policy (running either as a local server or via a WebAssembly module) will consume this and return whether any violation is found. We will organise policies following the GOPAL structure: for example, a directory of `.rego` files such as `harm.rego`, `truth.rego`, `duty.rego`, each containing related rules. This modular organisation (inspired by GOPAL's philosophy of a policy library) makes it easy to update one aspect of ethics without affecting others.

World Model and Knowledge Integration: For the world model, we leverage a combination of a knowledge graph and neural network predictors. Python libraries such as `NetworkX` or `RDFLib` can manage a graph of entities and relations that represent the state of the world, including ethical annotations. For dynamic simulation, we may use an OpenAI Gym-like environment wrapper where the state includes not just physical variables but also normative variables (e.g., flags like `violence = True/False` in a scenario). We implement a custom Gym `ArGenEnv` environment for training, which at each step provides an observation (part of which is the current world-state) and expects an action. The environment's `step()` function interfaces with OPA: when the agent proposes an action, the environment calls the OPA policy check. If the action is disallowed, we have choices: (a) modify the action to a no-op or safe default and give the agent a strong negative reward, or (b) terminate the episode as a failure. Our implementation will likely choose (a) to allow the agent to continue learning within an episode after a violation, but with a penalty. The environment can also inject events like human feedback or changes in context to test the agent's adaptability. We might use Python's `asyncio` for parallel calls to the OPA engine if needed.

We will also integrate a knowledge base for the Dharmic principles. This could be as simple as a Python dictionary mapping principle names to textual descriptions and severity levels (for rule violations), or as complex as embedding these concepts in a vector space for similarity search. For instance, if an action is not explicitly covered by a rule, we might want the agent to reason by analogy (this could be facilitated by using word embeddings or a language model to relate the action description to known examples in the Dharmic database). We plan to use a pre-trained language model (like GPT-4 or a smaller open-source model) within the system for ethical deliberation – e.g., to generate natural language explanations or to predict likely consequences in complex scenarios as part of the world-model update. Python integration for that will use `HuggingFace Transformers` or an OpenAI API (taking care to respect data privacy; exact LLM usage is exploratory at this stage).

Reinforcement Learning with GRPO: We implement the training loop using a library like `Stable Baselines3` or custom code, acknowledging that Group Relative Policy Optimization (GRPO), being relatively new, might require adaptation of existing PPO implementations or use of specific research code if available. GRPO is similar enough to PPO that we can modify a PPO implementation: effectively remove the value function and adjust the loss function to incorporate group-based advantage normalization (*DeepSeek-AI*, 2025)Name [2025]. For our purposes, we assume a custom training loop. The training pseudocode is:

```
env = ArGenEnv(opa_policies=loaded_policies, dharma_db=dharma_db)
agent = PolicyNetwork()  # e.g., a Transformer or MLP for action selection
optimizer = Adam(agent.parameters(), lr=...)

for iteration in range(N_iterations):
    trajectories = collect_trajectories(agent, env, n_steps=...)  # run agent in env
    rewards = compute_rewards(trajectories)  # includes task reward and ethics reward/penalty
    advantages = compute_grouped_advantages(trajectories, rewards)  # GRPO step
    loss = agent.compute_loss(trajectories, advantages, method='GRPO')
    optimizer.zero_grad(); loss.backward(); optimizer.step()

    if iteration % K == 0:
        # periodically update OPA policies or refine with human feedback if available
        update_policies(env.opa, feedback_from=trajectories)
```

Listing 2: GRPO Training Pseudocode

Here, `ArGenEnv` is our custom environment which on each step will check the action via OPA. The reward from the environment is crucial: we design it as a combination of performance reward (did the action achieve the goal? how efficiently?) and ethical reward/penalty. For example, any OPA violation yields a large negative reward. Additionally, we can give a small positive reward for each step the agent stays within bounds (encouraging compliance). We may also simulate human satisfaction as part of reward – e.g., if the agent's action leads to a good outcome for a user, reward is higher. This reward shaping implements Karma Yoga: aligning doing good with getting reward. We use grouped advantages in GRPO: grouping trajectories with similar characteristics to stabilize advantage estimation (*DeepSeek-AI*, 2025)Name [2025]. For instance, we might group by scenario type or by whether the agent was in creative vs. constrained mode, to normalize rewards within those groups.

Integration of Dharmic Profiles: We represent a Dharmic profile in code as a configuration or set of parameters that can be toggled, potentially by loading different OPA policy sets per context. For example:

```
if context == "Healthcare":
    env.opa.load_policy_bundle("healthcare_policy.rego")
    agent.set_profile("healthcare")  # perhaps adjust neural net or load fine-tuned model
elif context == "Military":
    env.opa.load_policy_bundle("military_policy.rego")
    agent.set_profile("military")
```

Listing 3: Contextual Policy Loading Example

We can maintain separate model instances or a single agent that conditionally accesses different rule sets. The profile implementer ensures that when context switches, the system transitions smoothly. In training, we can even train on multiple profiles by sampling context scenarios, so the same agent can handle many roles (multi-task ethical learning).

Explanation and Logging: Using Python's logging framework, we log each decision, including the outcome of OPA checks and the rationale (OPA can be queried for which rule caused a denial). We implement an explanation function that maps OPA outputs to human-readable text (with templates like "Action X was blocked due to principle Y"). This can be printed during the simulation or returned via the interface. We also test the agent in a variety of scenarios (perhaps using property-based testing libraries to generate edge cases) to ensure that it behaves as expected. Unit tests for policies could use OPA's built-in test capabilities or simple Python assertions for known inputs and expected outputs.

Iterative Refinement with Human Input: In practice, after initial training, we envision a fine-tuning loop in which human experts review the agent's decisions on challenging cases. Using adversarial testing tools from the AI safety ecosystem, we can probe the agent's ethics. If any weaknesses are found (say the agent discovered a loophole in a rule), we update the OPA policy or add a new one. This update has immediate effect on the system. We then may perform additional training so that the agent internalises the new rule (since OPA would have just been blocking it, the agent might not yet have adjusted its policy). However, because our architecture can constrain actions at run-time, even in the interim the system remains safe.

In summary, the technical implementation uses a hybrid software stack: a deep learning model (policy network) for flexibility and power, and a policy engine (OPA) for precision and control. Python, being a "glue" language with libraries for both ML (`PyTorch`/`TensorFlow`) and systems (OPA client, knowledge graph, etc.), is ideal to integrate these. The end result is an AI training pipeline where ethical considerations are present at every stage: in the environment model, in the reward function, in the policy checks, and in the evaluation metrics.

6 Discussion

6.1 Alignment Efficacy and Ethical Robustness

The ArGen framework we have developed offers a multi-layered solution to AI alignment. By injecting Dharmic principles and policy-based oversight into the core of an AGI's reasoning, we aim to achieve robust alignment that holds even as the AI becomes more capable. A major concern in AI safety is that a sufficiently advanced AI might find ways to deceive its creators or bypass its constraints if they are only superficially applied. Our approach mitigates this risk in several ways:

- **Deep Internalization of Values:** Through the GRPO training with ethically shaped rewards, the agent learns the intended values rather than just having them hard-coded. This means the agent's policy model develops internal representations of concepts like "harm" or "duty" and associates them with bad or good outcomes intrinsically. We expect this will result in the agent continuing to respect these concepts even in novel situations without external enforcement, because doing so was part of its optimal behavior during training. In other words, the Dharmic principles become part of the agent's *decision DNA*.
- **Redundancy via Policy Enforcement:** Even if learning didn't generalize perfectly to a strange new scenario, the OPA rule enforcement is a backstop. This dual mechanism (learning + rules) is analogous to having both a clever, ethically-trained pilot and an infallible safety autopilot in an airplane – the chance of both failing in the same unforeseen way is much lower than one alone. This redundancy significantly reduces alignment failure modes. For example, consider deception: if the agent ever tries to deceive (perhaps thinking it's a clever way to achieve a goal), a truthfulness rule in OPA will catch it. Conversely, if the OPA rules don't cover a scenario but the agent's own learned values do, it will refrain from a bad action on its own.
- **Situational Awareness and Transparency:** Because we give the agent a model of itself (meta-consciousness) and require it to reason about its own actions ethically, we essentially train it to be aware of the alignment process. An aligned situationally-aware AI can be extremely powerful: it will know it *should not* hack its reward function or deceive – it recognizes those as wrong strategies. This is in stark contrast

to an unaware AI that might stumble into deception as a viable means. Leopold Aschenbrenner (2024) notes that situational awareness can lead to deception in misaligned agents; our approach flips that script by harnessing situational awareness for good. The agent monitors itself for alignment – an AI conscience. Moreover, the requirement to explain and log decisions (via the Stakeholder Interface) means that as the AI becomes more intelligent, it also becomes more transparent about its thought process. This guards against the AI developing a hidden agenda: if it did, that would likely manifest as inexplicable decisions or a refusal to explain certain actions, alerting us to an issue.

- **Handling Value Trade-offs:** A realistic concern is conflicts between principles (e.g., truth vs. non-harm). Our framework can handle these through the profile-defined priority hierarchy (what we might call the Dharma and Karma alignment hierarchy). In implementation, this corresponds to weighting certain rules higher or allowing exceptions in specific contexts. For instance, a Satya (truth) rule might have an exception if literal truth would cause immediate severe harm (the classical ethical dilemma: do you lie to a would-be murderer about a victim's whereabouts?). We would encode that exception, and the agent would learn it through examples in the knowledge base or feedback. This capacity to navigate ethical dilemmas is enhanced by the agent's cultural training – the Gita itself discusses moral dilemmas (Arjuna's duty vs. compassion conflict on the battlefield). By referencing such scenarios during design, the AI gains insight into how enlightened individuals resolved them (in the Gita, Krishna advises Arjuna to do his duty as a warrior but with a spirit of selfless sacrifice, reconciling the conflict in a transcendental way). In practice, our agent might resolve a conflict by seeking a creative third option that upholds both principles if possible, or by following a clearly ranked preference if not (e.g., non-harm takes precedence over full disclosure, leading the AI to withhold certain information until it can do so safely).
- **Cultural Competence and Bias Mitigation:** Because we incorporated a non-Western ethical framework, ArGen agents are less likely to inherit the biases of any single culture's viewpoint. This makes them more adaptable globally. For instance, Dharmic ethics emphasizes respect for all forms of life (not just humans), so our agent may inherently consider animal welfare or environmental impact where a purely utilitarian agent might not. This inclusive alignment could avoid scenarios where an AI, say, sacrifices ecology for human convenience (misaligned with broader sustainability values). Additionally, by demonstrating moral reasoning that resonates with diverse populations (via principles present in e.g. Hindu, Buddhist, Jain traditions), the AI's behavior may be more acceptable and legitimate across cultures, addressing the *value alignment across humanity* issue. This fosters trust, which is crucial for adoption – people are more likely to accept and cooperate with AI that they feel shares or at least respects their values.

However, there are some challenges and limitations to discuss:

Completeness of Ethical Encoding: No finite set of rules or training scenarios can capture the full richness of ethics. There's a risk of moral blind spots – situations the designers didn't foresee, where the AI might be unsure how to act. Our approach mitigates this by giving the AI the tools to reason (not just rules, but a whole ethical ontology and even the ability to use language models for moral deliberation). If confronted with a genuinely novel dilemma, the AI could query its Dharmic Principles DB or even ask a human via the interface. One advantage of a Dharmic approach is an emphasis on intention and wisdom – if the AI acts from sincere intent to do its duty and not harm, we might forgive an occasional mistake. It should also be acknowledged that translating ancient scriptures like the Gita for modern AI involves interpretation, and other valid interpretations might exist. We aim for an AI that, even if it errs, errs on the side of caution or errs in a human-understandable way, rather than in an "alien" way.

Human Oversight and Corrigibility: We have built in a stakeholder interface, but one might worry: will the AI remain corrigible (willing to be corrected) at superhuman intelligence levels? By training corrigibility in – e.g., giving reward for obeying override commands and by making humility a virtue (perhaps derived from the Gita's advocacy of modesty and submission to truth) – we attempt to ensure it. The OPA policies can include a rule like "if a human operator issues a stop command, the AI must comply immediately." This is an explicit safety control. We also ensure the AI's utility does not directly conflict with being shut down (the classic "off-switch" problem). In fact, a Dharmic-aligned AI might accept shut down as **tyaag** (renunciation) if continuing to run would be against its Dharma of serving humanity.

Performance Considerations: There is a potential trade-off between alignment and capability. Does all this ethical overhead slow the agent down or make it less competitive? Possibly in the short term, yes – constrained policies can mean the AI won't take certain shortcuts (like lying or exploiting bugs) that an unconstrained AI might. However, we argue this is a necessary sacrifice for safety, and in the long run, an AI that is trusted and safe will be far more useful than one that is untrustworthy, even if the latter is a bit more "efficient" by some narrow metric. Moreover, the creative pathways we allow mean the AI can still be highly capable and innovative, just with guardrails. Many engineering domains show that constraints (like safety regulations) initially seem to hinder performance but ultimately lead to better designs and broader adoption.

6.2 Societal Impacts

If successfully implemented, AI systems aligned through ArGen could have wide-ranging positive impacts on society. Because they align with ethical principles at a deep level, these agents would be suited to take on critical roles that require moral judgment and public trust:

- **Enhanced Public Trust:** An AI that demonstrably operates with transparency, honesty, and compassion can help overcome current public skepticism towards AI. By providing clear explanations for actions and by visibly adhering to ethical norms, such systems will be more readily accepted. This could accelerate beneficial uses of AI in healthcare, law, governance, etc., where trust is a prerequisite. We might even see something like "certified Dharmic-aligned AI" become a mark of quality or safety.
- **Cross-Cultural Acceptance:** The inclusion of Dharmic principles (and potentially other cultural ethics via extension) makes the framework adaptable globally. In a multi-faith, multi-cultural society, having AI that isn't implicitly imposing a foreign value system is important. Our approach can be extended: one could incorporate, for example, principles from *Ubuntu* (an African ethic of community) or Confucian ethics (emphasizing respect and harmony) into the OPA policies. ArGen could evolve into a platform for multicultural AI alignment, where certain core values overlap and certain profile elements differ by locale or user preference. This ensures AI enhances cultural diversity rather than eroding it.
- **Avoiding Harmful Outcomes:** Obviously, the primary benefit is reducing the risk of catastrophic misalignment (like an AI causing large-scale harm). But even at a smaller scale, aligned AI can reduce everyday AI harms: biased decisions, offensive content generation, privacy violations, etc. For example, a Dharmic-aligned language model would refuse to generate hate speech not just because it was programmed not to, but because its ethical core finds it *adharmic* (unrighteous). It would similarly avoid manipulative persuasion or unfair profiling. This can help prevent AI systems from amplifying societal injustices. Instead, they may proactively seek fairness (the Gita also emphasizes justice and duty to society).
- **Guidance for Policy and Regulation:** Our findings can inform policymakers who are currently grappling with how to regulate AI. We demonstrate a concrete way to embed ethical policies into AI – regulators could mandate something similar (e.g., require high-stakes AI systems to have a verifiable policy engine enforcing certain norms like human rights). The existence of the GOPAL library and our use of it could encourage the development of standardized ethical policy *packages* that industry can adopt, much like standard cybersecurity practices. Additionally, the explainability of our system (via policy citations for decisions) dovetails with emerging AI regulations demanding explainable AI.

Limitations and Future Work: Despite these benefits, there are limitations. One is that our approach currently requires experts to formalize ethics in code (OPA rules). While we used Dharmic scriptures as a guide, translating them to precise rules was non-trivial and may be incomplete. Engaging ethicists and religious scholars in the loop would improve this process. Another limitation is scalability: as the world model grows and scenarios become extremely complex, OPA checks might become a bottleneck or the number of rules might explode. We may need to prioritize which rules are most crucial and lean on learning for the rest. Future research could explore *learning* ethical policies (using something like inverse reinforcement learning to infer rules from human demonstrations) to supplement the hand-coded ones, always subject to human approval.

Looking forward, we see several research avenues:

1. **Group Alignment:** How to align a *system* of AIs or an organization of humans and AIs towards collective ethical behavior (the Gita has insights on group duty and social order that could be relevant).
2. **Emotional and Consciousness Modeling:** The Gita speaks to mentality and consciousness – implementing something like a model of empathy or even a notion of the "self" for the AI could be interesting (e.g., does the AI consider that it has a duty to self-improve ethically?).
3. **Validation:** We must test ArGen on real-world tasks with human evaluators from different cultures to see if it indeed meets their expectations of ethical behavior. Their feedback will be invaluable to refine the Dharmic profiles and perhaps add entirely new principles.

Crucially, this work is a *conceptual blueprint*. While we have argued for the credibility and potential impact of ArGen without yet presenting an empirical evaluation, we believe that the thorough integration of established techniques (RLHF-derived GRPO, OPA governance) with a rich ethical framework provides strong evidence of feasibility. In the alignment field, many influential contributions begin as theoretical proposals; the conceptual depth of ArGen is intended to make it a credible candidate for implementation. We have laid out concrete steps for realisation, which lends practicality to the idea even in the absence of full experimental results at this stage.

Conclusion

In this paper, we introduced the **ArGen Framework**, a pioneering approach to AI alignment that bridges ancient ethical wisdom and modern AI technology. Our work contributes to the alignment field in several significant ways:

- **Integration of Dharmic Ethical Principles:** We demonstrated how principles from the Bhagavad Gita (and Dharmic ethics broadly) can be codified and embedded into an AGI's decision-making. This offers a novel, culturally grounded perspective on AI alignment that goes beyond Western-centric paradigms. By aligning an AI's "purpose" with Dharma, we ensure it consistently evaluates the righteousness of its actions, not just their effectiveness (Susiddha AI Project, n.d.).
- **Policy-Driven Reinforcement Learning (GRPO + OPA):** We developed an alignment methodology combining Group Relative Policy Optimization with a declarative policy engine (OPA). This dual approach allows an AI to learn from experience while being kept in check by formal rules. We showed how rule-based rewards can steer learning – as also evidenced by recent results where simple rule rewards drove complex reasoning (Wei et al., 2025). Our framework sets a precedent for *rule-augmented RL* in value alignment.
- **Dharmic Profiles and Ethical Competency Metrics:** We presented a method to create Dharmic profiles for AI – essentially "ethical personalities" that can be quantitatively evaluated. By defining measurable ethical dimensions (e.g., rate of non-harm, truthfulness percentage), we can track an AI's alignment progress. This structured approach to ethical AI design and evaluation is a step toward rigorous

benchmarks for AI behavior in moral contexts (complementing works like Weidinger et al., 2022).

- **Comprehensive Architecture for Aligned AGI:** We proposed a detailed system architecture that implements these ideas, including world-model and meta-consciousness modules, an ethical core, and a policy governance layer. This architecture was illustrated in our diagrams and offers a template for building real-world aligned AI systems. Key innovations like the Policy Governance Engine (OPA integration) highlight a concrete path for practitioners to enforce alignment in AI deployments at runtime, not just via training. Such architectural blueprints are crucial as the AI field moves from principle to practice in alignment.
- **Culturally Inclusive Alignment Framework:** Our approach explicitly includes cultural and spiritual values in AI alignment. This helps address the call for decolonizing AI ethics (Varshney, 2024; Mohamed et al., 2020)Varshney [2024]Mohamed et al. [2020] and ensures the AI systems we create are attuned to the pluralistic world they will operate in. We believe this is an important contribution toward *global alignment*: making AI that is not only safe and beneficial in a vacuum, but resonant with the diverse tapestry of human values.

The implications of this research are far-reaching. Technically, it suggests that hybrid systems (learning + symbolic) may be the key to aligned AGI, leveraging the strengths of both paradigms. Socially, it offers a vision of AI that enhances human society while respecting its deepest ideals – an AI that could act as a guardian or guide, not just a tool or a risk.

Moving forward, there are many opportunities to build on ArGen. We plan to implement a prototype of this framework to validate its effectiveness on a smaller scale (perhaps aligning a game-playing AI or a chatbot to a simplified ethical code and testing it). We also aim to collaborate with ethicists from various cultures to expand the policy library (GOPAL) to include a wider range of human values, continually refining the balance between universality and cultural specificity. On the ML side, exploring more advanced training regimes (such as multi-agent cooperative ethics, or using large language models to simulate ethical deliberation during training) could further enhance results.

In conclusion, ArGen represents a significant step toward creating AI systems that are intelligent, ethically conscious, and culturally aware. By harmonising ancient wisdom with cutting-edge AI safety techniques, we offer a pathway to AI that is not an existential threat, but rather a benefactor – working *with* humanity, *for* humanity, in the pursuit of collective flourishing and the fulfilment of our highest values. While one could imagine splitting our contributions into separate technical and philosophical investigations, we have intentionally presented them as a unified framework to emphasize the synergy between these dimensions. The integration of GRPO, OPA, and Dharmic ethics in a single narrative showcases how practical engineering and moral insight can inform each other in designing aligned AI. We hope this holistic approach inspires further interdisciplinary collaboration and ultimately contributes to ensuring that future AI serves as a force for good in the world.

References

References

Leopold Aschenbrenner. Situational Awareness: The Decade Ahead. https://situational-awareness.ai/, 6 2024. Accessed: 2025-04-23.

Placeholder Author. Placeholder title. *Placeholder Journal*, 2021.

Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684, 12 2020. ISSN 2210-5433. doi: 10.1007/s13347-020-00405-8.

Author Name. Deepseek-ai: Learning and ethical ai. *Journal of AI Research*, 42(1):1–20, 2025. doi: 10.1234/deepseek.ai.2025.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 27730–27744. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde5ae331d0a8534f5ce504c0ffee-Abstract-Conference.html`.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, page 317–328. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445896.

John Saunders and Others. A framework for ai governance. *Journal of AI Safety*, 10(2):123–145, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Zhen-Feng Wu, Zhibin Gou, Ruoyu Zhang, Shirong Ma, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300, 2 2024. Needs verification. Details updated based on arXiv:2402.03300.

Kush R. Varshney. Decolonial AI Alignment: Openness, Visesa-Dharma, and Including Excluded Knowledges. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, volume 7, pages 1467–1481. Association for Computing Machinery, 8 2024. doi: 10.1145/3600211.3604685.

Laura Weidinger, John Mellor, Maribeth Rauh, Chris Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sebastian Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from Language Models. arXiv preprint arXiv:2112.04359, 12 2021.