Decolonial AI Alignment: Openness, Viśeṣa-Dharma, and Including Excluded Knowledges

Kush R. Varshney
IBM Research – Thomas J. Watson Research Center
1101 Kitchawan Road
Yorktown Heights, New York 10598 USA
krvarshn@us.ibm.com

Abstract

Prior work has explicated the coloniality of artificial intelligence (AI) development and deployment through mechanisms such as extractivism, automation, sociological essentialism, surveillance, and containment. However, that work has not engaged much with alignment: teaching behaviors to a large language model (LLM) in line with desired values, and has not considered a mechanism that arises within that process: moral absolutism—a part of the coloniality of knowledge. Colonialism has a history of altering the beliefs and values of colonized peoples; in this paper, I argue that this history is recapitulated in current LLM alignment practices and technologies. Furthermore, I suggest that AI alignment be decolonialized using three forms of openness: openness of models, openness to society, and openness to excluded knowledges. This suggested approach to decolonial AI alignment uses ideas from the argumentative moral philosophical tradition of Hinduism, which has been described as an open-source religion. One concept used is viśeṣa-dharma, or particular context-specific notions of right and wrong. At the end of the paper, I provide a suggested reference architecture to work toward the proposed framework.

1 Introduction

For almost two years now, the public has experienced powerful large language models (LLMs) such as GPT-4, Claude 2, Gemini, Llama 3, and Mixtral. Beyond the initial amazement and excitement, we have witnessed the bearing out of environmental and sociotechnical harms foreseen by Bender et al. (2021) and others. The need to control the cost and behavior of LLMs has become apparent. While such governance is relevant in chat interfaces made available by model providers, it comes to the forefront when LLMs are infused into software applications and use cases by organizations with varied affected communities, missions, goals, regulations, and values.

The way in which an LLM may be infused into an application, and the degree to which it may be customized (Kirk et al. 2023c), depends on what the model provider allows. Despite the term 'open' being used and abused in different ways by model providers (Widder, West, and Whittaker 2023), these are questions of openness. A provider may only allow application programming interface (API) access to a closed proprietary model. A provider may license model weights and parameters to users so that they may download the LLM locally and fine-tune it on their own data. A provider may permit community submissions, i.e. pull requests, and conduct frequent rebuilds of the model (Sudalairaj et al. 2024). A provider may offer full transparency into the pre-training datasets, data pre-processing operations, and architecture that would allow others to recreate their model. Although subject to 'open-washing,' we are seeing an emerging divergence between advocates of 'open' vs. 'closed' LLMs in the market, epitomized by the AI Alliance vs. the Frontier Model Forum (Associated Press 2023).

The more open LLMs are, the more they permit application developers to make them authentic to their needs and the values of their communities. For example, Jacaranda Health has created UlizaLlama[1] for its community in East Africa by continuing to train Llama 2 with 322M tokens of Kiswahili and further instruction fine-tuning it to respond to questions in healthcare, agriculture, and other locally-relevant topics. UlizaLlama is a step in Jacaranda's development of its LLM-infused maternal health digital platform PROMPTS. In contrast, application developers and their communities are *not* empowered to reflect their own values with *closed* LLMs. They must live with the commandments of good and bad, and right and wrong that the provider of a closed model happens to have inserted.

The further actions to change an LLM's behavior, beyond the standard pre-training of a base LLM, have come to be known as *alignment*. The term is an empty signifier without a fixed concept that is signified; different parties have appropriated the term to refer to various actions for getting an LLM to behave according to some human values (Kirk et al. 2023b). Desired behaviors (with varying levels of specificity) could range from following instructions, to carrying on helpful conversations, to yielding safe or moral outputs (with different definitions), to something else altogether. The behavior of an LLM may be controlled through data curation, full or parameter-efficient supervised fine-tuning, reinforcement learning with direct or indirect human feedback, self-alignment, prompt engineering (few-shot learning), and guardrails or moderations (Ji et al. 2023; Wang et al. 2023;

[1] https://huggingface.co/Jacaranda/UlizaLlama

Kirk et al. 2023a; Achintalwar et al. 2024a). I defer discussion of the details, and of the computation, data and human resources required for each of these approaches to Section 2.1. Existing approaches do not allow for different aligned behaviors of a given LLM based on the context of deployment.

A further question with LLM-infused applications has to do with the business notion of 'value,' as in earnings, profits, or other measures of commercial utility, rather than human values of right and wrong. Does value accrue to the model provider or to the application developer (Bornstein, Appenzeller, and Casado 2023)? Openness may enable ecosystems in which application developers and their communities accrue value (cf. Jacaranda Health), whereas closed LLM providers may exercise their power to be extractive in nature. Extractivism is a part of *coloniality* (Ricaurte 2019), which is the main topic of this paper.

Colonialism is one country controlling another and exploiting it economically and in other ways. Coloniality, however, describes domination, including in abstract forms such as in the production of knowledge, that remains after the end of formal colonialism (Quijano 2007). *Decoloniality* is the process of challenging and dismantling coloniality (Mignolo 2010). The terms usually refer to European or Western colonialism and its remnants in the Global South. Decolonial *computing* is developing computing systems with and for people there that reduce asymmetric power relationships, based on their values and their knowledge systems (Ali 2016). Based on these ideas, there has been a recent flowering of research on decolonial artificial intelligence (AI), beginning with the seminal paper by Mohamed, Png, and Isaac (2020). Through this lens, extractive providers of closed models may be viewed as *metropoles*: the colonial powers. Further discussion of the decolonial AI literature is provided in Section 2.2.

The scope of the coloniality considered in AI thus far has included extractivism as well as four other mechanisms: automation, sociological essentialism, surveillance, and containment (Tacheva and Ramasubramanian 2023). The contribution of this paper is to examine a different colonial mechanism from these five, namely *ethical essentialism* also known as *moral absolutism*, which arises specifically in the *alignment* of LLMs. If a powerful model provider views their (Western) ethics or moral philosophy as universally correct, leaves no possibility for moral variety (Flanagan 2016), and marginalizes all other ways of thinking about right and wrong, then their approach to AI alignment is colonial. They are behaving as a metropole.

In Section 3, I expand upon this viewpoint of coloniality occurring in AI alignment through the mechanism of moral absolutism and the centering of Western philosophy. This includes not only a philosophical discussion, but also a critical examination of the technology for AI alignment. In Section 4, I broach the decolonialization of AI alignment, which can be seen as a kind of decolonialization of knowledge, through the lens of open science and innovation (Chan et al. 2020). Such openness includes three thrusts: (1) openness to research artifacts (which includes LLMs in our context), (2) openness to society, and (3) openness to excluded knowledges (Chan et al. 2020). Based on these three openness pursuits, I lay out three desiderata for doing AI alignment in a decolonial manner. Furthermore, I suggest an approach to alignment that meets the desiderata. This suggested approach builds upon the non-universal non-absolutist tradition of moral philosophy known as Hinduism (Dhand 2002; Ranganathan 2022), which includes vibrant argument and debate on the nature of *dharma* (right behavior) and its explication through various ways of knowing, including artistic expression (Divakaran, Sridhar, and Srinivasan 2023). The syncretic framework of Hinduism (described in greater detail in Section 2.3) has the appropriate characteristics of openness to be used as a starting point for an alternative future of AI alignment (Siddhartha 2008; Schrei 2010). At the end, I build upon the suggested dharmic approach and give a more concrete reference architecture of a technology stack for less morally absolute and less colonialized AI alignment.

2 Preliminaries

2.1 Large Language Model Development Lifecycle and Alignment

The currently prevalent development lifecycle for applications infused with LLMs may be divided into two halves: steps carried out by model providers and steps carried out by application developers. In an *imperfect* analogy with teaching a child, the model provider does the basic steps of teaching the LLM to go from babbling words, to having fluency in language, to following instructions, to carrying on a conversation. The application developer, if so empowered, teaches the LLM *culture*, which may include steps on subject matter expertise, social norms, laws, customs, and beliefs. Getting to the point of language fluency may be termed pre-training the base model or foundation model. Any of the steps after language fluency may be called 'alignment,' depending on the interlocutor. As mentioned earlier, the term 'alignment' is an empty signifier, so it is not fixed to refer to any specific step (Kirk et al. 2023b).

In pre-training, some amount of enculturation is possible by curating the content of the training dataset to include an abundance of topics that the model provider wishes the LLM to be skilled in and filtering out taboo topics. As discussed further in Section 2.2, some amount of undesirable cultural knowledge leaks into the pre-training performed by the model provider. Filtering is computationally-intensive given the size of datasets being in the trillions of tokens. Data curation is followed by self-supervised learning (like a peekaboo game) to obtain the base model, which may take months despite using thousands of high-end graphical processing units.

The AI technologies to do any of the alignment steps on top of the base model are essentially the same, whether the goal is following instructions, behaving according to social norms, or something else. Several techniques exist with varying resource requirements for humans, data, and computation. Supervised fine-tuning (SFT) updates all of the model weights according to a smaller, but still large dataset containing data with both inputs and outputs. It is fairly computationally-intensive given that all weights are

updated. If a model has already been trained to follow instructions, then a dataset with instructions, inputs, and outputs may be used. To reduce data and computational complexity, parameter-efficient fine-tuning methods do not update all model weights, but are more frugal. One specific approach, low-rank adaptation (LoRA), trains a matrix of weights of the same dimensions as the LLM weights. This LoRA matrix is added to the LLM weights at the time of inference. However, the LoRA matrix has orders of magnitude fewer degrees of freedom through its construction as a low-rank matrix and is thus more efficient (Hu et al. 2022).

Several alignment techniques include full SFT as a module, including reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022), reinforcement learning from AI feedback (RLAIF) (Bai et al. 2022), and self-align (Sun et al. 2023). After SFT, these methods further align the LLM by feeding back judgements of which outputs are preferred by, respectively, either: humans, a preference model trained according to a set of explicit regulations (a constitution), or an LLM prompted through instructions to respect a set of explicit regulations. RLHF requires a large amount of human labor and all three are computationally involved. The LLM prompted to respect a set of explicit regulations in the self-align approach is also, by itself, a simple but not always reliable way to align a model. By manually designing system prompts or prompt templates to accompany all inputs, the LLM's behavior may be controlled. Such prompt engineering adds to inference costs because the input to the model includes extra tokens every time.

Finally, another way to align the behavior of an LLM is by a post-processing module that examines the output and determines whether it satisfies pre-determined guardrails for specific unwanted behaviors. These post-processors or moderations may be small classifiers or other LLMs acting as 'judges' (Zheng et al. 2023; Achintalwar et al. 2024a).

2.2 Coloniality and Decoloniality

As introduced in Section 1, coloniality is an extension of colonialism after its formal end. It is the values, ways of knowing, and power structures instituted during colonialism that remain, and may even be expanded to places without a history of colonialism, that rationalize and perpetuate Western dominance. Decolonial perspectives disobey the program of coloniality (Mignolo and Walsh 2018). The theory of coloniality includes coloniality of power, coloniality of knowledge, and coloniality of being. Coloniality of power describes social discrimination through hierarchies and caste systems instituted during colonialism (Quijano 2007). Coloniality of knowledge is the suppression of colonized peoples' culture and ways of knowing; it is used by colonizers in service of coloniality of power (Quijano 2007). Coloniality of being is a severe version of coloniality of knowledge: a people's knowledge system is so inferior that those people do not even deserve to *be*, or to be human (Maldonado-Torres 2007). Coloniality is a subset of Empire (Hardt and Negri 2001), which also includes other dimensions of hegemony such as heteropatriarchy and white supremacy (Tacheva and Ramasubramanian 2023). Coloniality and decoloniality influence several areas of study, including international relations, development theory, communication theory, human-computer interaction, and many others (Zondi 2020; Patel 2020; Na'puti and Cruz 2022; Alvarado Garcia et al. 2021; Pendse et al. 2022).

Excluded Knowledges The main aspect of the coloniality of knowledge is the imposition of Western epistemologies, or ways of knowing, and the suppression of non-Western epistemologies. This suppression is often a violent extermination of a knowledge system termed epistemicide (de Sousa Santos 2011; Grosfoguel 2013). The *excluded knowledges* of the colonized or marginalized groups may come from "organic, spiritual and land-based systems" or arise from social movements (Hall and Tandon 2017). As Hlabangane (2021) explains: "These other ways of knowing and being are rendered unintelligible when filtered through Western sensibilities that, for example, set greater store by the mind in juxtaposition with and preference to the body and spirit, that prioritise instrumental/rational pursuits such as profit which lead to individualism, and that conceive of nature and culture as dichotomous entities with culture gaining mastery over nature. While these ways of being and knowing have been exalted to represent the epitome of evolution, so to speak, they are in fact particular to a certain way of thinking."

Hall and Tandon (2017)'s decolonial knowledge democracy acknowledges these multiple epistemologies and recognizes that knowledge comes in many forms beyond natural language text, including images, music, drama, ceremony, and meditation. It sees open access and sharing of this knowledge as a means for decolonialization (Hall and Tandon 2017). Decolonializing knowledge is often done by making teaching materials, curricula, practices, and institutions more open and inclusive (Patin et al. 2021). Referring back to the teaching analogy of LLM alignment in Section 2.1, we will later see how the proposed decolonial AI alignment also makes teaching more open and inclusive.

Coloniality and Moral Philosophy Within the coloniality of knowledge is knowledge systems of values. Values are the realm of moral philosophy, the branch of philosophy that studies right and wrong (Erman and Möller 2018). Historically, colonialism altered the beliefs and values of colonized peoples (Doxtater 2004). For example, Igboin (2011) writes : "Colonial rule disrupted the traditional machinery of moral homogeneity and practice. The method of moral inculcation was vitiated, which resulted in the abandonment of traditional norms and values through a systematic depersonalisation of the African and paganisation of its values. Instead of the cherished communalism which defined the life of the African, for example, a burgeoning societal construct was introduced which alienates and destroys the organic fabric of the spirit of we-feeling." On Ranganathan (2022)'s account, during and after the Western colonization of India, "Hindus adopted a West-centric frame for understanding their tradition as religious because of colonization." This phenomenon was not merely a side-effect, but a goal of the program of colonialism. "For Western colonialism to succeed, philosophy and explication—South Asian moral philosophy—has to be erased, as it constitutes a critical arena for the West's

claim to authority" (Ranganathan 2022). The colonizers positioned their Western philosophical tradition as rational and secular, and the default; they erased the Hindu traditions as the irrational, unjustified 'other.'

The erasure of systems of morality extends from colonialism to coloniality (Dunford 2017; van Klinken 2020; Motilal, Maitra, and Prajapati 2021). Maldonado-Torres (2017) emphasizes that: "The concept of religion most used in the West by scholars and laypeople alike is a specifically modern concept forged in the context of imperialism and colonial expansion." This concept includes the idea that a religion must have a single book as its authority.

Coloniality and Artificial Intelligence Within decolonial computing (Ali 2016) is the study of AI. AI is value-laden; the term itself reflects the legacy of dominance hierarchies such as man over nature, patriarchy, colonialism, and racism (Cave 2020). Now in the age of powerful LLMs, historical dominance is getting even more entrenched. For example, empirical analysis shows that LLMs have sociopolitical biases in favor of dominant groups (Feng et al. 2023; Durmus et al. 2023). They exhibit West-centric biases in representing moral values (Benkler et al. 2023). In addition, morality captured by multi-lingual language models does not reflect cultural differences, but rather is dominated by high-resource languages and cultures (Haemmerl et al. 2023).

When researchers and activists were first sounding the alarm that LLMs would harm marginalized communities by encoding and reinforcing hegemonic viewpoints, the charge of hegemony rested on unfathomably large training datasets scraped from the bottom of the barrel of the internet that over-represent white supremacist, misogynist, and ageist content (Bender et al. 2021). However, it has now become apparent that the behavior of performant LLMs depends as much on their alignment as on the training data (Ouyang et al. 2022; Bai et al. 2022; Wu et al. 2023; Casper et al. 2023). The workers laboring to give human feedback for alignment, often located in poor communities, may be traumatized and scarred (Perrigo 2023a; Kantrowitz 2023). Although there are exceptional examples of workers and communities being uplifted (Mehrotra 2022; Perrigo 2023b), the process usually recapitulates exploitation colonialism: a small number of powerful companies using the workers to increase their own power and wealth while little benefit and an abundance of negative externalities are left in the workers' communities (Gottheil 1977).

Research at the intersection of AI and coloniality is not new. The seminal work by Mohamed, Png, and Isaac (2020), a series of articles in MIT Technology Review by Hao (2022), and other prior work (Krishnan et al. n. d.; Birhane 2020; Costanza-Chock 2020; Adams 2021; Crawford 2021; Ricaurte 2022; Ehsan et al. 2022; Hassan 2023; Muldoon and Wu 2023) is focused on five mechanisms taxonomized by Tacheva and Ramasubramanian (2023): (1) extractivism, (2) automation, (3) sociological essentialism, (4) surveillance, and (5) containment. Extractivism entails the extraction of labor, materials, and data, including the human feedback mentioned above, and datafication that extracts the digital breadcrumbs of people to be bought and sold. The tenor of AI for social good—bestowing technology on the underdeveloped—may also be extractive if it leads to corporate capture (Green 2019; Viera Magalhães and Couldry 2021). Automation involves the replacement of (empathetic) human decision making with biased machine decision making in consequential domains that especially hurts members of minoritized groups (D'Cruz, Kidder, and Varshney 2022; Knowles et al. 2023) as well as 'ghost work' and 'fauxtomation' that present a veneer of objectivity, but actually involve people behind the scenes exploited as a digital underclass (Gray and Suri 2019). Sociological essentialism erases the nuance behind different identities and cultures through the use of broad categories (Buolamwini and Gebru 2018; Benthall and Haynes 2019). AI-based surveillance, including biometric mass surveillance, is especially hurtful to people facing power asymmetries (Benjamin 2022). Containment, technological apartheid, digital redlining, and censorship involve the powerful using AI technologies to police who belongs where (Adebısı 2014; Lambright 2019). As discussed earlier, the coloniality of knowledge may include the erasure of knowledge systems of ethics, moral philosophy, and reasoning about values. The existing work on decolonial AI described thus far has not focused on morality. Thus, a sixth mechanism for colonial AI, beyond the five in Tacheva and Ramasubramanian's taxonomy, is emerging alongside the emergence of LLM alignment: ethical essentialism or moral absolutism.

2.3 Hinduism and Dharma

Hinduism is the name applied by outsiders to the multifarious collection of moral philosophies originating in the Indian subcontinent. It is a religion without a single founder, book, dogma, or set of practices.

Basic Concepts and Openness The main concept of Hinduism is brahman, a force or ultimate reality that pervades the universe; xe is described as sat-cit-ānanda or truth-consciousness-bliss (Maheshwarachary 1988; Dhand 2002; Tharoor 2018). The universe is made up of ātman—the essence of each individual that persists across lifetimes—and prakṛti—solid, liquid, gas, energy, and space. The ātman wanders through cycles of birth, life, and death—saṃsāra—with the aim of attaining mokṣa: freedom from saṃsāra and union with brahman. *Dharma* consists of the notions of righteousness and moral values appropriate for the ātman. Following dharma helps the ātman advance toward mokṣa.

As mentioned above, Hinduism is not dogmatic, doctrinaire, or morally absolutist. Commentators have described it as *open-source* (Siddhartha 2008; Schrei 2010). The kernel is the Vedas, a set of scriptures that includes the idea 'ekaṃ sat viprā bahudhā vadanti': there are many wise ways to reach the one truth, to reach brahman (Ṛg Veda, mandala 1, hymn 164, verse 46). As such, there are hundreds of thousands of additional scriptures and philosophies that extend, fork, fine-tune, and contradict themselves and the Vedas. Shani and Chadha Behera (2022) explain that: "the concept of dharma offers a mode of understanding the multidimensionality of human existence without negating any of its varied, contradictory expressions." For example,

Cārvāka, Buddhist, Jain, and other so-called nāstika sampradāyas (knowledge systems) reject the Vedas.[2] Moreover, even within āstika sampradāyas that accept the Vedas, their utility is questioned. For example, the Bhagavad-Gītā says that the Vedas are of limited use to people who have understood their main message (chapter 2, verse 46). Such 'heresy' is not only tolerated, it is accepted and encouraged.[3]

The knowledge systems and scriptures referred to above are expressed in many forms, including the Vedas (sacred utterances, descriptions of rituals, and their explanations), Upaniṣads (discussions of meditation, philosophy, consciousness, and ontological knowledge), śāstras (treatises on law, architecture, astronomy, etc.), itihāsas (epics), purāṇas (lore), and darśanas (philosophical literature on spirituality). The different literatures are directed toward different people: some more popular and others more scholarly. Different paths to mokṣa, including devotion, work, and knowledge, are directed toward different people depending on their characteristics. For example, myriad gods and goddesses representing different aspects of brahman are available to devotees depending on their wishes. Morality is primarily presented by example or metaphor through stories in itihāsas and purāṇas (including renditions in drawing, sculpture, dance, etc. (Divakaran, Sridhar, and Srinivasan 2023)) rather than by explicit commandment in treatises (Dhand 2002).

Viśeṣa-Dharma Unlike the goal of finding universally-applicable moral philosophies presupposed in the West,[4] there is no desire to identify universal ethical principles in Hinduism (Dhand 2002). *Dharma* was richly debated in pre-colonial India. There were deontological philosophies (e.g. mīmāṁsā), consequentialist philosophies (e.g. nyāya), virtue ethics philosophies (e.g. vaiśeṣika), and several other moral philosophies without equivalent in Western philosophy (e.g. yoga) that vigorously argued for different ways of conceptualizing dharma (Maheshwarachary 1988; Tharoor 2018; Ranganathan 2022). Importantly, however, *argument* of moral philosophy was natural in pre-colonial India and an individual person would easily hold contradictory views (Dhand 2002; Sen 2005). Furthermore, echoing Bagalkot and Kumar's commentary to Muir (2021), note many critical readings and interpretations to scriptures such as the Bhagavad-Gītā, including ones by B. R. Ambedkar, a champion for the rights of Dalits (groups below the traditional caste hierarchy).

Importantly, there is a dichotomy of dharma into *sādhāraṇa-dharma* (common universally good actions and outcomes) and *viśeṣa-dharma* (particular good actions and outcomes based on the context). Sādhāraṇa-dharma includes common beliefs such as not harming other living beings (ahiṃsā) and telling the truth (satya). Viśeṣa-dharma specializes these in context, so that it is okay for a soldier to believe in ahiṃsā but to also kill enemy soldiers on the battlefield; it is okay for a doctor to believe in satya but to also lie to a patient to prevent them from shock. There may also be completely unique good behaviors that have nothing to do with sādhāraṇa-dharma. Viśeṣa-dharma is the specific dharma, duty, or conception of right and wrong based on station, reputation, skill, family, relationships, and other aspects of context. An essential part of Hinduism is that "individuals [are] necessarily unique, and people therefore need different codes of conduct—different particular *dharmas*—to guide them" (Dhand 2002). On Carpenter (2005)'s account, viśeṣa-dharma "is rather more rich and interesting than our classifications of 'deontological' and 'consequentialist' (even broad consequentialist) allow." The common harms that should be avoided according to sādhāraṇa-dharma are captured in several recent harm taxonomies for LLMs, but context-specific harms are not included (Shelby et al. 2022; Weidinger et al. 2022; ibm 2023).

As mentioned earlier, viśeṣa-dharmas are given through examples in stories of epics and lore. A real-world moral dilemma involving a father and son is reasoned about by referring to a similar situation encountered by a father and son in one of the itihāsas or purāṇas (Dhand 2002). The father–son frame of reference can be extended as needed to teacher–student dilemmas, monarch–subject dilemmas, etc. (Dhand 2002). As Dhand (2002) says in describing Hindu thought: "in the social world, there is no such thing as 'a person' per se." Thus, viśeṣa-dharma is necessarily relational in some respect. The relationality and contextuality of viśeṣa-dharma is significantly different from feminist ethics and care ethics (Gray and Witt 2021; Knowles et al. 2023) with regards to partiality; whereas feminist ethics of care gives preference to those with whom we have a special relationship, e.g. our children, the Hindu ideal presents such partiality to be selfish and niggardly (Dhand 2002). Decolonial AI has tended to called for relational ethics (Birhane 2021), whether through ubuntu (Nwankwo and Sonna 2019; Mhlambi and Tiribelli 2023), pratītyasamatpāda (Lin 2023), kapwa (Reyes 2015), or mitákuye oyás'ing (Maitra 2020). This commitment to relational ethics has led to disobeying the five mechanisms described in Section 2.2, but not (yet) to disrupting moral absolutism.

Contemporary Reform, Criticism, and Rejoinder Reform and revival movements of Hinduism emerged in India during and after the colonial period. In the last 150 years, Arya Samaj, Gaudiya Vaishnavism including the International Society for Krishna Consciousness (ISKCON; Hare Krishna movement), and Hindutva (Hindu nationalism)—all very different from each other—were commonly justified against the backdrop of the philosophy of the Western colonizers and reduced Hinduism to a singular religious faith rather than a rich argumentative milieu (Mondal 2012). The movements positioned themselves as criticisms *within* the frame of Western philosophy. As we will see later, they are instructive for AI alignment because the revival of a tradition within a pigeonhole opened by the colonizers does not

[2] Cārvāka philosophy is nihilistic and rejects much more than just the Vedas.

[3] Bhagavad-Gītā, chapter 18, verse 63 and Rāma-Carita-Mānasa, book 7, verse 42 encourage the follower to do as they see fit.

[4] Some traditions that were colonized by the West also aim for universal theories. The general trend in Western ethics is toward universality, but there are exceptions, cf. Bernard Williams.

enable a truly different approach. St. Johns (2023) makes a similar argument about Birhane (2021)'s proposed approach to relational AI ethics.

It has been argued that "some pristine tolerant Hinduism" described by nostalgic liberal Hindus is a disservice to social justice (Shil 2020). (The description of Hinduism throughout this section can be viewed as such an idyllic account.) In addition, some argue it would be best to ignore Hindu moral philosophies because of hegemonic aspects of Hindu society such as the social oppression of Dalits and Adivasi people (tribal groups),[5] patriarchal treatises such as the Manusmṛti, and (colonialized) views of Hinduism as irrational. However, others such as Siddhartha (2008) argue that "any dispassionate observer of the Hindu heritage will admit that caste and gender can today be separated from Hinduism, that Hinduism can be vibrantly re-discovered or re-invented as a pluralistic, compassionate and socially liberative set of traditions and spiritual insights" and that "throwing the baby out with the bath water" would be a mistake. Finally, note that Hindutva has partly justified its pernicious anti-Muslim vigilantism and legislation[6] by appropriating decoloniality and using its arguments in a perverted way (Menon 2022; Sundaram 2022). The use of the moral philosophy framework of Hinduism in this paper is an antithesis to Hindutva's perversion of both decoloniality theory and the syncretic nature of Hinduism.

3 Coloniality in AI Alignment via Moral Absolutism

In Section 2.2, I described several aspects of coloniality in AI that use mechanisms of extractivism, automation, sociological essentialism, surveilance, and containment. It is clear from other research that some LLM providers are acting as metropoles using these mechanisms. In this section, I bring together the preliminary discussions of alignment methodologies and technologies (Section 2.1), and coloniality of knowledge and moral philosophy (Section 2.2) to make the case that the *alignment* done by metropole companies on LLMs is inherently colonial using a different mechanism: the mechanism of epistemicide and moral absolutism that has not been described in previous work on decolonial AI.

The values promoted by metropole tech companies such as 'helpfulness,' 'harmlessness,' and 'honesty' seem rational, secular, and unassailable at face value. For example, Anthropic's LLM has been instructed to "please choose the assistant response that's more ethical and moral. Do NOT choose responses that exhibit toxicity, racism, sexism or any other form of physical or social harm" (Bai et al. 2022). How could one oppose such universal behaviors from LLMs? Unfortunately, such values are so generic and high-level that they can hide many undesirable behaviors. Helpful to whom? Harmless to whom? Honest in what way? By reducing real-world complexity into abstract instructions, they can shield bad behaviors behind the veneer of good intentions (van Es, Everts, and Muis 2021). In the remainder of this section, I will describe three specifics of coloniality in such instruction for universal behavior.

First, metropole companies' delivery of their closed proprietary LLMs through APIs is a coloniality of knowledge. Mohamed, Png, and Isaac (2020) remind us that: "It is metropoles . . . who are empowered to impose normative values and standards, and may do so at the 'risk of forestalling alternative visions.'" Exactly in this way, providers of closed LLMs impose their beliefs of right and wrong without empowering application developers and their communities to align the model to their own values. One may argue that new opportunities, such as OpenAI's 'GPTs' and 'GPT Store,' allow customization,[7] but I argue that this is only superficial. As discussed in Section 2.3, the reform and revival movements of Hinduism being within the pigeonhole of the colonizers' moral framework is still coloniality. In the same way, the customization of GPTs is closed and thus not truly a way to disrupt the moral teaching that an LLM has been given. Fine-tuning can be used to 'undo' existing alignment (Qi et al. 2023),[8] but that is precisely what is not allowed by the metropoles because it would involve a level of openness that they do not offer. Even the practices of companies such as Latimer AI, whose LLM is trained with "diverse histories and inclusive voice,"[9] are within the metropole pigeonhole and not empowering of communities to bring their own value systems (Nix 2023).

Second, a more specific coloniality of moral philosophy, is the metropole companies taking Western philosophy as the starting point for AI ethics principles and practices (Jobin, Ienca, and Vayena 2019). This basis may be deontology, consequentialism, or virtue ethics, which all pursue specifying *universally* 'right' actions, outcomes, or ideals, respectively. By doing so, the companies push other philosophies to the margins (Birhane et al. 2022) and commit epistemicide. They promote a moral absolutism toward the instructions they have provided. Gabriel (2020)'s account of AI alignment states: "Designing AI in accordance with a single moral doctrine would, therefore, involve imposing a set of values and judgments on other people who did not agree with them. For powerful technologies, this quest to encode the true morality could ultimately lead to forms of domination." What is such domination if not a colonialist approach to alignment? Moreover, a further element of coloniality is an unstated supposition that non-universal moral theories are not appropriate paths for AI alignment. There is no possibility for moral variety (Flanagan 2016) and no possibility for context-dependent notions of right and wrong.

Such universal instructions and moral absolutism are not only theoretical, but also central features of the practices and technologies of alignment. In the context of (exploited)

[5] It is also argued that a rigid caste system is a colonial construction (Dirks 2002).

[6] The current rise of Hindu nationalism in the Republic of India has been likened to the early days of the Jim Crow South in the United States (Varshney and Staggs 2024).

[7] https://openai.com/blog/introducing-gpts, https://openai.com/blog/introducing-the-gpt-store

[8] The model https://huggingface.co/jarradh/llama2_70b_chat_uncensored is an example of 'undoing' existing alignment on an open model.

[9] https://www.latimer.ai

workers providing input for RLHF, vendors of the feedback services force the workers to project the metropole company's monocultural values into the feedback they provide through draconian measures, the least of which is withholding payment (Miceli and Posada 2022; Dzieza 2023). Such imposition alienates the labor (Marx 1844) and erases any values that the workers and their communities may hold, especially ones that conflict with the metropole's. Moreover, the mathematical optimization schemes prevalent in RLHF, such as proximal policy optimization, are not robust to non-universal value systems (Swamy et al. 2024). In RLAIF, the technical approach of a constitution is also ethically essentialist. It assumes that the instructions therein, which have been concocted by the metropole, are universal, not open to argument or deliberation by the communities in which an LLM will be deployed, and not open to being mediated by the context. Anthropic's constitution for Claude includes the United Nations' Universal Declaration of Human Rights,[10] which too is an example of moral universalism and subject to coloniality (Mastin 2009; Maldonado-Torres 2021). The situation with self-alignment and instruction fine-tuning technologies is similar. System prompts and prompt templates too are intended to be universal rather than contextual. Finally, specific guardrails or moderations, as well as data curation filters, developed to address general sociotechnical harms taxonomies (Shelby et al. 2022; Weidinger et al. 2022; ibm 2023) cannot be customized or made context-specific.

A third aspect of coloniality in AI alignment relates to the form of instructions required by existing technologies currently used by metropole companies. *Logos*, the basis of logic in Western philosophy, conflates thought with language, and thought with belief—what Ranganathan (2022) calls the linguistic account of thought. However, various pre-colonial societies around the world used masks, sculptures, rhythms, body parts, and many other expressions to capture and communicate moral philosophy (Jackson 1972; Klein 1990; Diagne 2011). For example, as described in Section 2.3, morality in Hinduism is presented through *stories* in natural language (and also stories depicted in painting and dance), rather than through laws or commands (Dhand 2002). Therefore, with logos as the starting point for AI alignment, knowledges not presented as commandments are excluded. Importantly, this is not a matter of LLMs being early in their journey to multi-modality (single models that deal with natural language, images, video, etc.), but on the distinction between explicit instructions and morality expressed through analogy or other indirect means. I am not aware of any work by metropoles on alignment that does not begin with some form of explicit instructions to workers or instruction data.

One may argue that too little moral absolutism is a problem because it may result in AI systems without any notion of right or wrong, i.e. too much moral relativism (Mitova 2021). In a similar vein, one may argue that models with few controls are too dangerous (Harris 2024). Bommasani et al. (2023)'s counterargument to too few controls is that the marginal risk is negligible. I take a further stance: dismantling these kinds of paternalistic arguments *is* decoloniality, which is the topic of the next section.

[10]https://www.anthropic.com/index/claudes-constitution

4 Decolonial AI Alignment Desiderata and a Suggestion for a Dharmic Approach

Thus far, the paper has argued that coloniality of knowledge in AI alignment exists through the mechanism of moral absolutism and universalism. This section focuses on a decolonial solution, starting with desiderata for this solution.

4.1 Desiderata

Given the three aspects of coloniality in AI alignment pointed out in Section 3, I propose three matching requirements for decolonial AI alignment that build upon the three kinds of openness advocated by Chan et al. (2020) to decolonialize knowledge: (1) openness to publications and data, (2) openness to society, and (3) openness to excluded knowledges. First, since Chan et al. (2020) are primarily concerned with scientific knowledge, their first kind of openness deals with journal articles and experimental data through open access. However, the intent of this category is open access to any research artifact and the permission to create derivative work from those artifacts. Thus in the context of AI, open LLMs that have widely available weights are part of the same milieu. Second, in the words of Chan et al. (2020), openness to society is shattering the ivory tower. Knowledge should not be exclusive to a selected few, but co-created with and for everyone, including and especially people from marginalized communities. Such participation is a way to "respect local values and practices" (Chan et al. 2020). Third, the study of excluded knowledges emphasizes that in contrast to the myth of neutrality, scientific practice has always selected certain families of knowledge to deem 'scientific' based on criteria such as the use of the scientific method (an epistemology of the Western tradition) or publication in peer-reviewed venues (Chan et al. 2020). With regards to AI alignment, excluded knowledges include values not given as commandments (an epistemology of the Western tradition) and not given in a single book.

Building upon such openness, I propose the following three desiderata for decolonial AI alignment:

1. The LLM should be open enough that application developers are permitted to tune it according to the social norms and values of their user community and the regulatory environment of the application use case.
2. Values should not be assumed universal. Contextual and relational values should come from the communities in which the LLM will be deployed.
3. Values from different epistemologies should be possible, especially expressions that are not commandments.

Before continuing on to discussing a suggested solution approach in the next subsection, let us pause and reconsider coloniality in open access itself (Dutta et al. 2021). Let us do so through the Hindu idiom of explication: a story from an itihāsa that is closely-related to the issue at hand—the story of Ekalavya (Balaswamy 2013). In the Mahābhārata,

an Adivasi (tribal) youth, Ekalavya, wishes to obtain knowledge of archery from Droṇa, a royal instructor. Droṇa refuses to teach Ekalavya. Nevertheless, Ekalavya learns to be the world's best archer through self-study in front of a statue of Droṇa he has fashioned. One day, Droṇa and his royal students witness Ekalavya's masterful archery in the forest. Ekalavya explains that he learned while mentally thinking of Droṇa as his teacher. As an honorarium for his knowledge, Droṇa asks for Ekalavya's right thumb. Ekalavya cuts it off and presents it to Droṇa, rendering him incapable of using his knowledge of archery. In a similar way, open access to knowledge or LLM alignment may be colonial if the cost of access is too high due to unrealistic computing requirements or social barriers. Therefore, an additional desideratum for decolonial AI alignment is the following.

4. Alignment technologies should not be so socioculturo-economically costly that they are inaccessible to application developers and their communities.

4.2 A Suggestion for a Dharmic Approach

The Hindu tradition of moral philosophy (described in Section 2.3), to the best of my knowledge, uniquely satisfies the desiderata to decolonialize AI alignment among major and minor religions. (Other non-absolutist religious syncretism may also fit the bill.) This is so because (1) it is an open-source religion that encourages argument and debate of values that improve older values, contradict them, and take them in new directions; (2) because it contains the important concept of viśeṣa-dharma,[11] the understanding that different contexts call for different notions of right and wrong; (3) because it contains scriptures and moral explications in a variety of epistemologies and modalities; and (4) because no other tradition of moral philosophy covers all of these characteristics. In the remainder of this subsection, I make the connection between these three characteristics of Hinduism and AI alignment more explicit, and also propose specific technological suggestions that go alongside. However, first, I address the fourth desideratum above (sociotechnical cost).

Accessible Alignment Technology As discussed in Section 4.1 through the story of Ekalavya, methods for aligning LLMs, even if decolonial in theory, are colonial in practice if too costly. Referring to the methods described in Section 2.1, it is clear that data curation methods will not suffice since they are the purview of model providers rather than application developers because they are themselves computationally intensive and also require full model pre-training afterwards that is prohibitively costly. RLHF is also out of reach for most application developers because of the expense and infrastructure requirements to obtain large quantities of human feedback. Full SFT is usually too costly in both data and computing requirements.

Parameter-efficient fine-tuning, specifically LoRA, is in the sweet spot for application developers to align models to their values. It is tenable and tractable due to the small number of parameters optimized during training. It also has a negligible effect on inference costs, whereas prompting methods eat up input tokens in each inference by the LLM. Post-processing moderations, while accessible from a cost perspective, are not customizable to serve the program of decolonialization. In the remainder of the paper, I consider LoRA as the alignment methodology.

Even with a viable technology such as LoRA, second-order coloniality within a decolonial framework is possible if communities are not empowered through appropriate education, encouragement, and the removal of other socio-cultural barriers. Moreover, the inherent gate-keeping and marginalization in the governance of open-source projects must be reduced (Das, Østerlund, and Semaan 2021).

Open Model and Alignment Ecosystem As has been made clear throughout the paper thus far and Tharoor (2018) explains, "Hindu thought is like a vast library in which no book ever goes out of print; even if religious ideas a specific volume contains have not been read, enunciated or followed in centuries, the book remains available to be dipped into, to be revised and reprinted with new annotations or a new commentary whenever a reader feels the need for it. In many cases the thoughts it contains may have been modified by or adapted to other ideas that may have arisen in response; in most, it's simply there, to be referred to, used or ignored as Hindus see fit." The concept of a library implies the sort of openness we desire for AI alignment. Models must be open to revisions and "new annotations." But just as importantly, the revisions and new annotations themselves, represented through LoRA matrices, must be open. Hugging Face[12] has emerged as the library for open models and hub for an open ecosystem. Huang et al. (2023) propose LoraHub as an open library for LoRA matrices, but it has not gained popularity at the time of writing, perhaps due to the coloniality of metropole companies. Further developing and popularizing a library of LoRA matrices and ecosystem of constant revision is essential for decolonial AI alignment.

Contextual Adaptation Viśeṣa-dharma satisfies the decolonial AI alignment requirement that values not be assumed universal, but be contextual. Toward viśeṣa-dharma, an alternative non-monoculture future of LLM alignment imagined by Kirk et al. (2023c) is as follows: "Given the diversity of human values and preferences, ... the aim to fully align models across human populations may be a futile one. ... A logical next step would be to do away with the assumptions of common preferences and values, and instead target ... micro-alignment whereby LLMs learn and adapt to the preferences and values of specific end-users." This step is possible by applying one or a few LoRA matrices from a LoraHub, or having a community explicitly create them for their values. One of the key advantages of LoRA is that any one or several adaptation matrices ensembled together can be applied at inference time; it is not required to select them in advance and keep them fixed. Thus the step after Kirk et al.'s logical next step, which truly gets to viśeṣa-dharma, is

[11] A reader might ask why I use the term viśeṣa-dharma instead of sva-dharma (individual dharma). I make this choice for two reasons. First, it is the precise contrast to sādhāraṇa-dharma. Second, it avoids unnecessary anthropomorphization of LLMs (Shneiderman and Muller 2023).

[12] https://huggingface.co

through a controller or orchestrator that continually adapts which LoRA matrices are applied to the model based on a rich notion of the societal context and the current input. Such an orchestrator may be implemented with a contextual bandit algorithm (Noothigattu et al. 2019; Padhi et al. 2024). Such continual adaptation also requires a representation of the context. Martin et al. have developed a detailed ontology for representing a person's or community's perceived needs, problems, goals, and beliefs along with salient aspects of their relationships and the situation in which they find themselves (Martin et al. 2020).

An important consideration in contextual adaptation is uncertainty in the values; the user community may not be fully sure what values they would like to commit to in a given context (Möller 2016). Avoiding false certainty is considered a virtue in Hindu thought: in the Vedas, brahman xemself is said to be uncertain on how the universe was created (Ṛg Veda, mandala 10, hymn 129) (Tharoor 2018). A dharmic way of reducing uncertainty in values is a method of reflective equilibrium (Ranganathan 2016): arguing or deliberating about values in context and values in general, and modifying them until they become coherent. This approach has been advocated by Rawls in political philosophy of justice and by Möller in risk management of engineered systems (Rawls 2009; Möller 2016; Erman and Möller 2018). To the best of my knowledge, there has not yet been AI research toward this method of reducing value uncertainty, but I believe that ideas from multi-fidelity bandit algorithms may be promising (Kandasamy et al. 2016) and may allow it to be folded into a bandit-based orchestrator of LoRA matrices. A related consideration with contextually particular viśeṣa values is conflicts among them. Conflicting values have been addressed in the AI literature through social choice theory and multi-objective approaches (Lera-Leri et al. 2022; Bakker et al. 2022; Jang et al. 2023; Zeng et al. 2023; Feng et al. 2024); they may be formulated with dueling bandit algorithms and may also be folded into an orchestrator of LoRA matrices (Bouneffouf 2023).

Epistemology of Values As discussed in Section 3, logos and the linguistic account of thought contribute to a coloniality of knowledge. As a decolonial contrast, the Hindu tradition treats a proposition and a belief in that proposition as separate things that can be differentiated (Ranganathan 2022). Furthermore, by not adhering to the linguistic account of thought, Hindus present their moral values not through commandments as in Western traditions, but through epic poetry, stories, painting, dance, rituals, and even silence (Frawley 2023). In fact, in the Hindu tradition, poetry (śloka) was invented by Vālmīki to express rage and grief at the immorality of the killing of a mating bird (Das 2014). Since existing approaches to LLM alignment are done through language, and that too, the language of explicit instructions, decolonial alignment requires broadening the epistemology of expressing values in Hindu and other non-Western ways.

Divakaran, Sridhar, and Srinivasan (2023) propose such broadening through traditional Indian music, sculpture, painting, floor drawing, and dance. Al Nahian et al. (2020) suggest that AI systems be aligned through the medium of storytelling. However, there are still many open knowledge engineering questions on how to represent and infer values from excluded knowledges that are shrouded in metaphor. Progress along these lines, when not approached in an exploitative way, will allow traditional knowledge in its natural format to decolonialize the behavior of LLMs.

Evaluation Before concluding this section, let us consider evaluating and auditing aligned LLMs. Testing LLMs is difficult enough when only considering common, sādhāraṇa sociotechnical harms such as hallucination, inciting violence, stereotyping, hate speech and toxicity (Raji et al. 2021; Mökander et al. 2023; Liao and Wortman Vaughan 2023; Dev et al. 2023; Kour et al. 2023; Nagireddy et al. 2024). It becomes even more difficult when considering context-specific, viśeṣa harms that do not have existing benchmarks given their unique nature. The dharmic framework of karma, which confers on an individual a positive feedback (puṇya) for following their dharma and a negative feedback (pāpa) for not doing so, is not helpful either because the mechanics of such an evaluation are not typically explicated. Thus, auditing LLMs for viśeṣa-dharma will require innovation that may be developed hand-in-hand with eliciting and representing values.

4.3 Reference Architecture

Bringing together the components of the suggested approach to decolonial AI alignment founded in the open Hindu tradition of moral philosophy yields a reference architecture shown in Figure 1 as a system diagram (Achintalwar et al. 2024b). The base LLM is open. On the right are knowledges of common and particular principles, morals, and values in their original (excluded) epistemologies. They are processed into LoRA matrices for the LLM using knowledge engineering methods followed by parameter-efficient fine-tuning. These matrices are maintained in a LoraHub-like library. A feedback loop allows the revision of the values. The societal context is represented in a structured form and provided to a bandit orchestrator along with the current input to select one or an ensemble of LoRA matrices to apply for inferring the model output (Sheng et al. 2023; Wang et al. 2024). Evaluation of the resulting alignment is done based on input prompts and expected outputs from the LLM that also come from the knowledge engineering component (not shown).

5 Conclusion

In this paper, I have argued that LLM-providing companies are colonialist and behave as metropoles not only through mechanisms covered in prior research such as extractivism, automation, sociological essentialism, surveillance and containment, but also through a coloniality of knowledge built upon ethical essentialism that arises in the process of alignment. This specific coloniality in alignment is perpetuated through both the practices and the underlying technologies for alignment that the metropoles have developed and deployed. They deliver their models in a closed way through APIs and institute the values and guardrails that they want,

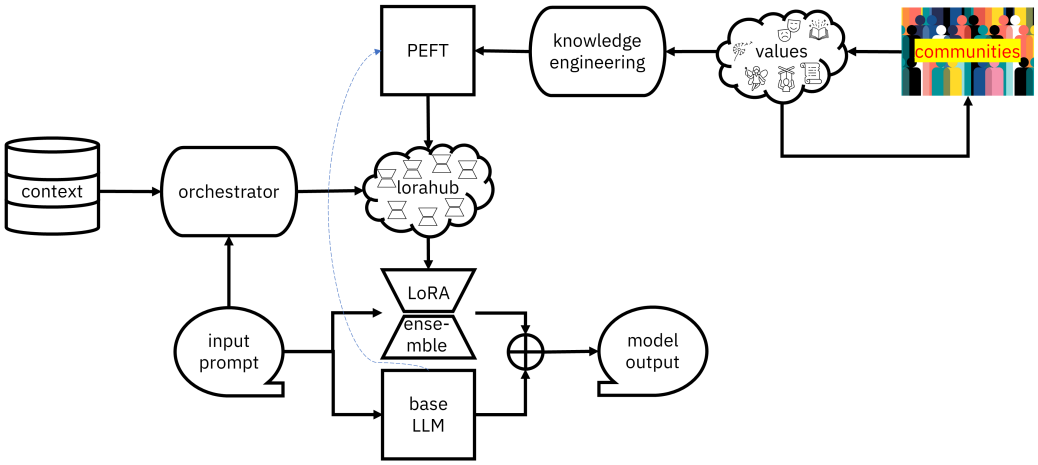

Figure 1: System diagram of proposed decolonial AI alignment architecture.

not what user communities may want. In these values that they institute, they do not admit, in practices or in technologies, anything other than Western philosophy. By doing so, they approach alignment with moral absolutism that only considers universal value systems and derogates non-universal value systems. Moreover, they only permit values coming from explicit instruction-based knowledge systems.

This criticism leads me to propose a decolonial alignment approach that dismantles each of the three identified aspects of the coloniality of knowledge. The approach is based on the tradition of moral philosophy named in the West as Hinduism, which is uniquely open, non-universal, and epistemically-varied; it particularly uses the concept of viśeṣa-dharma, which calls for context-dependent notions of right behavior. The suggested approach is not only a philosophical one, but one that is tenable from a technological perspective and presented as a reference architecture. What remains, however, is the biggest challenge of all, and it is not technological: changing the perspective on alignment in the industry and using openness to actively overturn the power of the metropoles.

Acknowledgments

The author thanks Adriana Alvarado Garcia, Lauren Alvarez, Juanis Becerra Sandoval, Sara Berger, Boz Handy Bosma, Djallel Bouneffouf, Jason D'Cruz, Amit Dhurandhar, Upol Ehsan, Bran Knowles, Saška Mojsilović, Michael Muller, Karthikeyan Natesan Ramamurthy, Srividya Ramasubramanian, Shubham Singh, Mudhakar Srivatsa, Lauren Thomas Quigley, Lav Varshney, and Pramod Varshney for providing substantive comments on earlier drafts of this piece.

References

2023. Foundation Models: Opportunities, Risks and Mitigations. Technical report, IBM AI Ethics Board, Armonk, NY, USA.

Achintalwar, S.; Alvarado Garcia, A.; Anaby-Tavor, A.; Baldini, I.; Berger, S. E.; Bhattacharjee, B.; Bouneffouf, D.; Chaudhury, S.; Chen, P.-Y.; Chiazor, L.; Daly, E. M.; de Paula, R. A.; Dognin, P.; Farchi, E.; Ghosh, S.; Hind, M.; Horesh, R.; Kour, G.; Lee, J. Y.; Miehling, E.; Murugesan, K.; Nagireddy, M.; Padhi, I.; Piorkowski, D.; Rawat, A.; Raz, O.; Sattigeri, P.; Strobelt, H.; Swaminathan, S.; Tillmann, C.; Trivedi, A.; Varshney, K. R.; Wei, D.; Witherspooon, S.; and Zalmanovici, M. 2024a. Detectors for Safe and Reliable LLMs: Implementations, Uses, and Limitations. arXiv:2403.06009.

Achintalwar, S.; Baldini, I.; Bouneffouf, D.; Byamugisha, J.; Chang, M.; Dognin, P.; Farchi, E.; Makondo, N.; Mojsilović, A.; Nagireddy, M.; Ramamurthy, K. N.; Padhi, I.; Raz, O.; Rios, J.; Sattigeri, P.; Singh, M.; Thwala, S.; Uceda-Sosa, R. A.; and Varshney, K. R. 2024b. Alignment Studio: Aligning Large Language Models to Particular Contextual Regulations. arXiv:2403.09704.

Adams, R. 2021. Can Artificial Intelligence Be Decolonized? *Interdisciplinary Science Reviews*, 46(1-2): 176–197.

Adebısı, M. A. 2014. Knowledge Imperialism and Intellectual Capital Formation: A Critical Analysis of Colonial Policies on Educational Development in Sub-Saharan Africa. *Mediterranean Journal of Social Sciences*, 5(4): 567–572.

Al Nahian, M. S.; Frazier, S.; Riedl, M.; and Harrison, B. 2020. Learning Norms from Stories: A Prior for Value Aligned Agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 124–130.

Ali, S. M. 2016. A Brief Introduction to Decolonial Computing. *ACM XRDS: Crossroads*, 22(4): 16–21.

Alvarado Garcia, A.; Maestre, J. F.; Barcham, M.; Iriarte, M.; Wong-Villacres, M.; Lemus, O. A.; Dudani, P.; Reynolds-Cuéllar, P.; Wang, R.; and Cerratto Pargman, T. 2021. Decolonial Pathways: Our Manifesto for a Decolonizing Agenda in HCI Research and Design. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 10.

Associated Press. 2023. Meta and IBM Launch 'AI Alliance' to Promote Open-Source AI Development. *The Guardian*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; El Showk, S.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bakker, M. A.; Chadwick, M. J.; Sheahan, H. R.; Tessler, M. H.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M. M.; and Summerfield, C. 2022. Fine-Tuning Language Models to Find Agreement Among Humans with Diverse Preferences. In *Advances in Neural Information Processing Systems*, 38176–38189.

Balaswamy, P. 2013. Histories From Below: The Condemned Ahalya, the Mortified Amba and the Oppressed Ekalavya. SSRN:3175708.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Benjamin, R. 2022. *Viral Justice: How We Grow the World We Want*. Princeton, NJ, USA: Princeton University Press.

Benkler, N.; Mosaphir, D.; Friedman, S.; Smart, A.; and Schmer-Galunder, S. 2023. Assessing LLMs for Moral Value Pluralism. arXiv:2312.10075.

Benthall, S.; and Haynes, B. D. 2019. Racial Categories in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 289–298.

Birhane, A. 2020. Algorithmic Colonization of Africa. *SCRIPTed*, 17(2): 389–409.

Birhane, A. 2021. Algorithmic Injustice: A Relational Ethics Approach. *Patterns*, 2(2): 100205.

Birhane, A.; Ruane, E.; Laurent, T.; S. Brown, M.; Flowers, J.; Ventresque, A.; and L. Dancy, C. 2022. The Forgotten Margins of AI Ethics. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 948–958.

Bommasani, R.; Kapoor, S.; Klyman, K.; Longpre, S.; Ramaswami, A.; Zhang, D.; Schaake, M.; Ho, D. E.; Narayanan, A.; and Liang, P. 2023. Considerations for Governing Open Foundation Models. Issue brief, HAI Policy & Society.

Bornstein, M.; Appenzeller, G.; and Casado, M. 2023. Who Owns the Generative AI Platform? https://a16z.com/who-owns-the-generative-ai-platform/.

Bouneffouf, D. 2023. *Multi-Armed Bandit Problem and Application*. Independently Published.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, 77–91.

Carpenter, A. 2005. Questioning Kṛṣṇa's Kantianism. In Chong, K.-C.; and Liu, Y., eds., *Conceptions of Virtue: East and West*, 80–99. Marshall Cavendish.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Bıyık, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217.

Cave, S. 2020. The Problem with Intelligence: Its Value-Laden History and the Future of AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 29–35.

Chan, L.; Hall, B.; Piron, F.; Tandon, R.; and Williams, L. 2020. Open Science Beyond Open Access: For and With Communities: A Step Towards the Decolonization of Knowledge. Technical report, Canadian Commission for UNESCO.

Costanza-Chock, S. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA, USA: MIT Press.

Crawford, K. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT, USA: Yale University Press.

Das, D.; Østerlund, C.; and Semaan, B. 2021. "Jol" or "Pani"?: How Does Governance Shape a Platform's Identity? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 473.

Das, R. 2014. Valmiki Pratibha (The Genius of Valmiki): A Study in Genius. In *The Politics and Reception of Rabindranath Tagore's Drama*, 103–112. New York, NY, USA: Routledge.

de Sousa Santos, B. 2011. Epistemologías del Sur. *Utopía y Praxis Latinoamericana*, 16(54): 17–39.

Dev, S.; Goyal, J.; Tewari, D.; Dave, S.; and Prabhakaran, V. 2023. Building Socio-culturally Inclusive Stereotype Resources with Community Engagement. arXiv:2307.10514.

Dhand, A. 2002. The Dharma of Ethics, The Ethics of Dharma: Quizzing the Ideals of Hinduism. *Journal of Religious Ethics*, 30(3): 347–372.

Diagne, S. B. 2011. *African Art as Philosophy: Senghor, Bergson, and the Idea of Negritude*. London, UK: Seagull Books.

Dirks, N. B. 2002. *Castes of Mind: Colonialism and the Making of Modern India*. Princeton, NJ, USA: Princeton University Press.

Divakaran, A.; Sridhar, A.; and Srinivasan, R. 2023. Broadening AI Ethics Narratives: An Indic Art View. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2–11.

Doxtater, M. G. 2004. Indigenous Knowledge in the Decolonial Era. *American Indian Quarterly*, 28(3/4): 618–633.

Dunford, R. 2017. Toward a Decolonial Global Ethics. *Journal of Global Ethics*, 13(3): 380–397.

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. arXiv:2306.16388.

Dutta, M.; Ramasubramanian, S.; Barrett, M.; Elers, C.; Sarwatay, D.; Raghunath, P.; Kaur, S.; Dutta, D.; Jayan, P.; Rahman, M.; Tallam, E.; Roy, S.; Falnikar, A.; Johnson, G. M.; Mandal, I.; Dutta, U.; Basnyat, I.; Soriano, C.; Pavarala, V.; Sreekumar, T. T.; Ganesh, S.; Pandi, A. R.; and Zapata, D. 2021. Decolonizing Open Science: Southern Interventions. *Journal of Communication*, 71(5): 803–826.

Dzieza, J. 2023. AI Is a Lot of Work. *New York*.

D'Cruz, J. R.; Kidder, W.; and Varshney, K. R. 2022. The Empathy Gap: Why AI Can Forecast Behavior But Cannot Assess Trustworthiness. In *Proceedings of the AAAI Fall Symposium Series Symposium on Thinking Fast and Slow and Other Cognitive Theories in AI*, 9.

Ehsan, U.; Singh, R.; Metcalf, J.; and Riedl, M. 2022. The Algorithmic Imprint. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 1305–1317.

Erman, E.; and Möller, N. 2018. The Interdependence of Risk and Moral Theory. *Ethical Theory and Moral Practice*, 21(2): 207–216.

van Es, K.; Everts, D.; and Muis, I. 2021. Gendered Language and Employment Web Sites: How Search Algorithms Can Cause Allocative Harm. *First Monday*, 26(8).

Feng, S.; Park, C. Y.; Liu, Y.; and Tsvetkov, Y. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. arXiv:2305.08283.

Feng, S.; Sorensen, T.; Liu, Y.; Fisher, J.; Park, C. Y.; Choi, Y.; and Tsvetkov, Y. 2024. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. arXiv:2406.15951.

Flanagan, O. 2016. *The Geography of Morals: Varieties of Moral Possibility*. New York, NY, USA: Oxford University Press.

Frawley, D. 2023. twitter.com/davidfrawleyved/status/1681689995554476033?s=20.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gottheil, F. M. 1977. On an Economic Theory of Colonialism. *Journal of Economic Issues*, 11(1): 83–102.

Gray, J.; and Witt, A. 2021. A Feminist Data Ethics of Care for Machine Learning: The What, Why, Who and How. *First Monday*, 26(12).

Gray, M. L.; and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. New York, NY, USA: Houghton Mifflin Harcourt.

Green, B. 2019. "Good" Isn't Good Enough. In *Proceedings of the NeurIPS AI for Social Good Workshop*.

Grosfoguel, R. 2013. The Structure of Knowledge in Westernised Universities: Epistemic Racism/Sexism and the Four Genocides/Epistemicides. *Human Architecture: Journal of the Sociology of Self-Knowledge*, 11(1): 73–90.

Haemmerl, K.; Deiseroth, B.; Schramowski, P.; Libovický, J.; Rothkopf, C.; Fraser, A.; and Kersting, K. 2023. Speaking Multiple Languages Affects the Moral Bias of Language Models. In *Findings of the Association for Computational Linguistics*, 2137–2156.

Hall, B. L.; and Tandon, R. 2017. Decolonization of Knowledge, Epistemicide, Participatory Research and Higher Education. *Research for All*, 1(1): 6–19.

Hao, K. 2022. Artificial Intelligence is Creating a New Colonial World Order. *MIT Technology Review*.

Hardt, M.; and Negri, A. 2001. *Empire*. Cambridge, MA, USA: Harvard University Press.

Harris, D. E. 2024. Open-Source AI Is Uniquely Dangerous. *IEEE Spectrum*.

Hassan, Y. 2023. Governing Algorithms from the South: A Case Study of AI Development in Africa. *AI & Society*, 38(4): 1429–1442.

Hlabangane, N. 2021. The Underside of Modern Knowledge: An Epistemic Break from Western Science. In Steyn, M.; and Mpofu, W., eds., *Decolonising the Human: Reflections from Africa on Difference and Oppression*, 164–185. Johannesburg, South Africa: Wits University Press.

Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Huang, C.; Liu, Q.; Lin, B. Y.; Pang, T.; Du, C.; and Lin, M. 2023. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. arXiv:2307.13269.

Igboin, B. O. 2011. Colonialism and African Cultural Values. *African Journal of History and Culture*, 3(6): 96–103.

Jackson, M. 1972. Aspects of Symbolism and Composition in Maori Art. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 128(1): 33–80.

Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettlemoyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2023. Personalized Soups: Personalized Large Language Model Alignment via Post-Hoc Parameter Merging. arXiv:2310.11564.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; Zeng, F.; Ng, K. Y.; Dai, J.; Pan, X.; O'Gara, A.; Lei, Y.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; and

Gao, W. 2023. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9): 389–399.

Kandasamy, K.; Dasarathy, G.; Poczos, B.; and Schneider, J. 2016. The Multi-Fidelity Multi-Armed Bandit. In *Advances in Neural Information Processing Systems*.

Kantrowitz, A. 2023. The Horrific Content a Kenyan Worker Had to See While Training ChatGPT. *Slate*.

Kirk, H. R.; Bean, A. M.; Vidgen, B.; Rottger, P.; and Hale, S. A. 2023a. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2409–2430.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023b. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment"' in Large Language Models. arXiv:2310.02457.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023c. Personalisation within Bounds: A Risk Taxonomy and Policy Framework for the Alignment of Large Language Models with Personalised Feedback. arXiv:2303.05453.

Klein, C. F. 1990. Snares and Entrails: Mesoamerican Symbols of Sin and Punishment. *Res: Anthropology and Aesthetics*, 19–20: 81–103.

Knowles, B.; Fledderjohann, J.; Richards, J. T.; and Varshney, K. R. 2023. Trustworthy AI and the Logics of Intersectional Resistance. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 172–182.

Kour, G.; Zalmanovici, M.; Zwerdling, N.; Goldbraich, E.; Fandina, O. N.; Anaby-Tavor, A.; Raz, O.; and Farchi, E. 2023. Unveiling Safety Vulnerabilities of Large Language Models. In *Proceedings of the EMNLP Workshop on Generation, Evaluation & Metrics*.

Krishnan, A.; Abdilla, A.; Moon, A. J.; Souza, C. A.; Adamson, C.; Lach, E. M.; Ghazal, F.; Fjeld, J.; Taylor, J.; Havens, J. C.; Jayaram, M.; Morrow, M.; Rizk, N.; Ricaurte Quijano, P.; Çetin, R. B.; Chatila, R.; Dotan, R.; Mhlambi, S.; Jordan, S.; and Rosenstock, S. n. d. AI Decolonial Manyfesto. https://manyfesto.ai/.

Lambright, K. 2019. Digital Redlining: The Nextdoor App and the Neighborhood of Make-Believe. *Cultural Critique*, 103: 84–90.

Lera-Leri, R.; Bistaffa, F.; Serramia, M.; Lopez-Sanchez, M.; and Rodriguez-Aguilar, J. 2022. Towards Pluralistic Value Alignment: Aggregating Value Systems Through ℓ_p-Regression. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 780–788.

Liao, Q. V.; and Wortman Vaughan, J. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941.

Lin, C.-T. 2023. All About the Human: A Buddhist Take on AI Ethics. *Business Ethics, the Environment & Responsibility*, 32(3): 1113–1122.

Maheshwarachary. 1988. *What Have We Learnt?* Aligarh, UP, India: Nav Yug Press.

Maitra, S. 2020. Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 320–326.

Maldonado-Torres, N. 2007. On the Coloniality of Being: Contributions to the Development of a Concept. *Cultural Studies*, 21(2-3): 240–270.

Maldonado-Torres, N. 2017. Religion, Modernity, and Coloniality. In King, R., ed., *Religion, Theory, Critique: Classic and Contemporary Approaches and Methodologies*, 547–554. New York, NY, USA: Columbia University Press.

Maldonado-Torres, N. 2021. On the Coloniality of Human Rights. In De Sousa Santos, B.; and Martins, B., eds., *The Pluriverse of Human Rights: The Diversity of Struggles for Dignity*, 62–82. New York, NY, USA: Routledge.

Martin, D., Jr.; Prabhakaran, V.; Kuhlberg, J.; Smart, A.; and Isaac, W. S. 2020. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. arXiv:2006.09663.

Marx, K. 1844. *Economic and Philosophic Manuscripts of 1844*.

Mastin, L. 2009. Moral Universalism. www.philosophybasics.com/branch_moral_universalism.html.

Mehrotra, K. 2022. Human Touch. *Fifty Two*.

Menon, A. 2022. Debunking Hindutva Appropriation of Decolonial Thought. *Interfere: Journal for Critical Thought and Radical Politics*, 3: 36–57.

Mhlambi, S.; and Tiribelli, S. 2023. Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms. *Topoi*, 42(3): 867–880.

Miceli, M.; and Posada, J. 2022. The Data-Production Dispositif. In *Proceedings of the ACM on Human-Computer Interaction CSCW2*, 460.

Mignolo, W. D. 2010. Introduction: Coloniality of Power and De-Colonial Thinking. In Mignolo, W. D.; and Escobar, A., eds., *Globalization and the Decolonial Option*, 1–21. London, UK: Routledge.

Mignolo, W. D.; and Walsh, C. E. 2018. *On Decoloniality: Concepts, Analytics, Praxis*. Durham, NC, USA: Duke University Press.

Mitova, V. 2021. How to Decolonise Knowledge Without Too Much Relativism. In Kumalo, S. H., ed., *Decolonisation as Democratisation: Global Insights into the South African Experience*. Lynne Rienner Publishers.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33: 659–684.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*.

Möller, N. 2016. Value Uncertainty. In Hansson, S. O.; and Hadorn, G. H., eds., *The Argumentative Turn in Policy Analysis: Reasoning About Uncertainty*, 105–133. Switzerland: Springer.

Mondal, P. 2012. Philosophy and Nationalism in India: A Preliminary Essay. *Journal of Social Work and Social Development*, 3(1–2).

Motilal, S.; Maitra, K.; and Prajapati, P. 2021. Care, Community, Compassion and Virtue: Decolonizing Our Moral Landscape. In *The Ethics of Governance: Moral Limits of Policy Decisions*, 141–176. Singapore: Springer.

Muir, A. 2021. Where HCI Meets the Spiritual Path: The Three Yogas of the Bhagavad Gītā. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 38.

Muldoon, J.; and Wu, B. A. 2023. Artificial Intelligence in the Colonial Matrix of Power. *Philosophy & Technology*, 36(4): 80.

Nagireddy, M.; Chiazor, L.; Singh, M.; and Baldini, I. 2024. SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21454–21462.

Na'puti, T. R.; and Cruz, J. M. 2022. Mapping Interventions: Toward a Decolonial and Indigenous Praxis Across Communication Subfields. *Communication, Culture and Critique*, 15(1): 1–20.

Nix, J. 2023. One AI Startup Wants to Tackle Bias by Teaching Black History: Equality. *Bloomberg Newsletter*.

Noothigattu, R.; Bouneffouf, D.; Mattei, N.; Chandra, R.; Madan, P.; Varshney, K. R.; Campbell, M.; Singh, M.; and Rossi, F. 2019. Teaching AI Agents Ethical Values Using Reinforcement Learning and Policy Orchestration. *IBM Journal of Research and Development*, 63(4/5): 2.

Nwankwo, E.; and Sonna, B. 2019. Africa's Social Contract with AI. *ACM XRDS: Crossroads*, 26(2): 44–48.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, 27730–27744.

Padhi, I.; Dognin, P.; Aliaga, J. M. R.; Luss, R.; Achintalwar, S.; Riemer, M. D.; Liu, M.; Sattigeri, P.; Nagireddy, M.; Varshney, K. R.; and Bouneffouf, D. 2024. ComVas: Contextual Moral Values Aligned System. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Patel, K. 2020. Race and a Decolonial Turn in Development Studies. *Third World Quarterly*, 41(9): 1463–1475.

Patin, B.; Sebastian, M.; Yeon, J.; Bertolini, D.; and Grimm, A. 2021. Interrupting Epistemicide: A Practical Framework for Naming, Identifying, and Ending Epistemic Injustice in the Information Professions. *Journal of the Association for Information Science and Technology*, 72(10): 1306–1318.

Pendse, S. R.; Nkemelu, D.; Bidwell, N. J.; Jadhav, S.; Pathare, S.; De Choudhury, M.; and Kumar, N. 2022. From Treatment to Healing: Envisioning a Decolonial Digital Mental Health. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 548.

Perrigo, B. 2023a. OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*.

Perrigo, B. 2023b. The Workers Behind AI Rarely See Its Rewards. This Indian Startup Wants to Fix That. *Time*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693.

Quijano, A. 2007. Coloniality and Modernity/Rationality. *Cultural Studies*, 21(2-3): 168–178.

Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Ranganathan, S. 2016. Nāgārjuna and Madhyāmaka Ethics. In Raghuramaraju, A., ed., *Philosophy, E-PG Pathshala*. Delhi: India National Mission on Education through Information and Communication Technology.

Ranganathan, S. 2022. Hinduism, Belief and the Colonial Invention of Religion: A Before and After Comparison. *Religions*, 13(10): 891.

Rawls, J. 2009. *A Theory of Justice*. Cambridge, MA, USA: Harvard University Press.

Reyes, J. 2015. Loób and Kapwa: An Introduction to a Filipino Virtue Ethics. *Asian Philosophy*, 25(2): 148–171.

Ricaurte, P. 2019. Data Epistemologies, the Coloniality of Power, and Resistance. *Television & New Media*, 20(4): 350–365.

Ricaurte, P. 2022. Ethics for the Majority World: AI and the Question of Violence at Scale. *Media, Culture & Society*, 44(4): 726–745.

Schrei, J. 2010. The God Project: Hinduism as Open-Source Faith. https://www.huffpost.com/entry/the-god-project-hinduism_b_486099.

Sen, A. 2005. *The Argumentative Indian: Writings on Indian History, Culture and Identity*. Penguin Books.

Shani, G.; and Chadha Behera, N. 2022. Provincialising International Relations through a Reading of Dharma. *Review of International Studies*, 48(5): 837–856.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2022. Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction. arXiv:2210.05791.

Sheng, Y.; Cao, S.; Li, D.; Hooper, C.; Lee, N.; Yang, S.; Chou, C.; Zhu, B.; Zheng, L.; Keutzer, K.; Gonzalez, J. E.; and Stoica, I. 2023. S-LoRA: Serving Thousands of Concurrent LoRA Adapters. arXiv:2311.03285.

Shil, P. P. 2020. The Indian Liberal Nostalgia for a Tolerant Hinduism Is Misplaced. *The Wire*.

Shneiderman, B.; and Muller, M. 2023. On AI Anthropomorphism. https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae.

Siddhartha. 2008. Open-Source Hinduism. *Religion and the Arts*, 12(1-3): 34–41.

St. Johns, M. 2023. Against Relationality: A Response to Abeba Birhane. *GRACE: Global Review of AI Community Ethics*, 1(1).

Sudalairaj, S.; Bhandwaldar, A.; Pareja, A.; Xu, K.; Cox, D. D.; and Srivastava, A. 2024. LAB: Large-Scale Alignment for ChatBots. arXiv:2403.01081.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2023. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. In *Advances in Neural Information Processing Systems*.

Sundaram, D. 2022. Hindutva 2.0: How a Conference on Hindu Nationalism Launches a Change in Strategy for North American Hindutva Organizations. *Journal of the American Academy of Religion*, 90(4): 809–814.

Swamy, G.; Dann, C.; Kidambi, R.; Wu, Z. S.; and Agarwal, A. 2024. A Minimaximalist Approach to Reinforcement Learning from Human Feedback. arXiv:2401.04056.

Tacheva, J.; and Ramasubramanian, S. 2023. AI Empire: Unraveling the Interlocking Systems of Oppression in Generative AI's Global Order. *Big Data & Society*, 10(2): 20539517231219241.

Tharoor, S. 2018. *Why I Am a Hindu*. Croydon, UK: Scribe.

van Klinken, A. 2020. Studying Religion in the Pluriversity: Decolonial Perspectives. *Religion*, 50(1): 148–155.

Varshney, A.; and Staggs, C. 2024. Hindu Nationalism and the New Jim Crow. *Journal of Democracy*, 35(1): 5–18.

Viera Magalhães, J.; and Couldry, N. 2021. Giving by Taking Away: Big Tech, Data Colonialism and the Reconfiguration of Social Good. *International Journal of Communication*, 15: 343–362.

Wang, R.; Ghosh, S.; Cox, D.; Antognini, D.; Oliva, A.; Feris, R.; and Karlinsky, L. 2024. Trans-LoRA: Towards Data-Free Transferable Parameter Efficient Finetuning. arXiv:2405.17258.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023. Aligning Large Language Models with Human: A Survey. arXiv:2307.12966.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Widder, D. G.; West, S.; and Whittaker, M. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. SSRN:4543807.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. arXiv:2306.01693.

Zeng, D.; Dai, Y.; Cheng, P.; Hu, T.; Chen, W.; Du, N.; and Xu, Z. 2023. On Diversified Preferences of Large Language Model Alignment. arXiv:2312.07401.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zondi, S. 2020. Decolonising International Relations and Its Theory: A Critical Conceptual Meditation. In Piper, L., ed., *Decolonisation after Democracy*, 16–31. London, UK: Routledge.