IMPERIAL

महाजनो येन गतः स पन्थाः

The path the great have walked; that is the path to follow. Mahābhārata, Vana Parva 3.313.117 (Yakṣa Praśna)

Dharma, Demonstration and the Alignment of Machines

Sonali Parbhoo

Imperial College London

Hinduism and AI Roundtable, University of Cambridge

September 17, 2026

Thanks to **Priyanka Parbhoo**, **Rupal Vora**, **Chinmaya Mission** and discussions with the AI4AI lab.

STANFORD LAW SCHOOL | AI INITIATIVE

Centre for Human Inspired Artificial Intelligence

BAPS SWAMINARAYAN RESEARCH

future of life INSTITUTE

AI4AI AI for Actionable Impact

धर्म

dharma

One word with several different applications



One word with several different applications

धर्म

dharma

सामान्य धर्म

sāmānya dharma

What holds for everyone

One word with several different applications

धर्म
dharma

सामान्य धर्म

sāmānya dharma

What holds for everyone

विशेष धर्म

viśeṣa dharma

What holds particularly

One word with several different applications

धर्म
dharma

सामान्य धर्म

sāmānya dharma

What holds for everyone

विशेष धर्म

viśeṣa dharma

What holds particularly

आपद्धर्म

āpad-dharma

What holds in exceptional situations

One word with several different applications

धर्म
dharma

कुलधर्म
kuladharma

राजधर्म
rājadharma

आश्रम धर्म
āśrama dharma

वर्णाश्रम धर्म
varṇāśrama dharma

स्वधर्म
svadharma

युगधर्म
yugadharma

सामान्य धर्म
sāmānya dharma

What holds for everyone

विशेष धर्म
viśeṣa dharma

What holds particularly

आपद्धर्म
āpad-dharma

What holds in exceptional situations

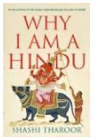
Can one system hold many dharmāḥ?

एकं सद् विप्रा बहुधा वदन्ति।

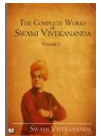
ekam sat viprāḥ bahudhā vadanti

Truth is one; the learned call it by different names – Ṛgveda 1.164.46c

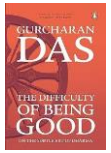
Hindu Thought



- **Plurality** – Hinduism accommodates multiple paths to dharma*.



- **Differences matter** – Different people and temperaments may require different paths**.



- **Dharma requires discernment** – Dharma can involve difficult or conflicting judgments^.
- **One deeper orientation** – Dharma serves a common spiritual goal.

*Shashi Tharoor. *Why I am a Hindu*. Oxford University Press, 2018.
**Swami Vivekananda. *The complete works of Swami Vivekananda*. Advaita Ashram, 1907.
***Neil Lawrence. *The Atomic Human: Understanding Ourselves in the Age of AI*, Penguin UK, 2024.
^ Gurcharan Das. The *Difficulty of being good*. Penguin UK, 2009

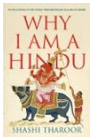
Can one system hold many dharmāḥ?

एकं सद् विप्रा बहुधा वदन्ति।

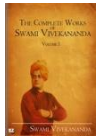
ekam sat viprāḥ bahudhā vadanti

Truth is one; the learned call it by different names – Ṛgveda 1.164.46c

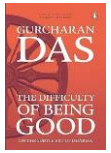
Hindu Thought



- **Plurality** – Hinduism accommodates multiple paths to dharma*.



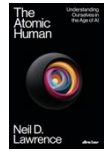
- **Differences matter** – Different people and temperaments may require different paths**.



- **Dharma requires discernment** – Dharma can involve difficult or conflicting judgments^.
- **One deeper orientation** – Dharma serves a common spiritual goal.

AI Alignment

- **Plurality** – Can one ML system represent multiple normative frameworks rather than encoding a single fixed objective?
- **Context matters** – Human intelligence is situated in culture, relationships and experience. Can behaviour adapt to the *recipient and context*?***



- **Alignment requires judgement** – Can we determine what considerations should govern a situation and when and how these apply, instead of *what people prefer*.
- Can an ML system maintain **stable higher-level commitments** while adapting its behaviour?

*Shashi Tharoor. *Why I am a Hindu*. Oxford University Press, 2018.
**Swami Vivekananda. *The complete works of Swami Vivekananda*. Advaita Ashram, 1907.
***Neil Lawrence. *The Atomic Human: Understanding Ourselves in the Age of AI*, Penguin UK, 2024.
^ Gurcharan Das. *The Difficulty of being good*. Penguin UK, 2009



Image Source: https://vediccosmos.com/yaksaprasnah-continued/



महाजनो येन गतः स पन्थाः

The path the great have walked; that is the path to follow.

Mahābhārata, Vana Parva 3.313.117 (Yakṣa Praśna)

Image Source: https://vediccosmos.com/yaksaprasnah-continued/

What does dharma require in practice?

Yudhiṣṭhira's answer to the Yakṣa serves as an example of dharma in practice. Mahābhārata, Vana Parva 313.117

The verse	What ML does now	The challenges
तर्कोऽप्रतिष्ठः *tarko 'pratiṣṭhaḥ* 3.316.117 Reasoning has no firm footing	**Constrained optimization.** Write down an objective to train a model and hard constraints, then optimize. *Achiam et al. 2017; Dai et al. 2024*	**Misspecified objectives.** In ML we frequently optimize on proxies for what we actually want through enforcing e.g. constraints on expected cost. What if this objective is incorrect? *Amodei et al. 2016; Gao et al. 2023*
श्रुतयो विभिन्ना *śrutayo vibhinnā* The scriptures are diverse	**Constitutional guidance.** Train against written principles; deliberative alignment reasons over a specification. *Bai et al. 2022; Guan et al. 2024*	**Disagreement in Principles**. In Constitutional AI a principle is sampled at random at each critique step, so conflicts are never adjudicated. What if principles conflict?
नैको ऋषिर्यस्य मतं प्रमाणम् *naiko ṛṣir yasya mataṃ pramāṇam* No single sage is authoritative	**Optimizing preferences.** Current ML fits one model to aggregated pairwise preferences. *Christiano et al. 2017; Ouyang et al. 2022*	One model averages away disagreement and reflects whoever was sampled. Pluralistic alignment **models the distribution**, not the mean. Whose judgement is represented? *Santurkar et al. 2023; Casper et al. 2023; Sorensen et al. 2024*

Governance: who decides, and when to step back

तदेतत्क्षत्रस्य क्षत्रं यद्धर्मः। तस्माद्धर्मात्परं नास्ति।

Dharma is the rule over the ruler; there is nothing higher than dharma. Bṛhadāraṇyaka Upaniṣad 1.4.14 (Kāṇva)

राजधर्म *rājadharma*

The ruler administers dharma, and is also bound by it.
Mahābhārata, Śāntiparvan

Human-in-the-loop

Place a person at critical decision points who can monitor, override or halt the system. High-risk AI now legally requires this.

EU AI Act, Art. 14; Green 2022

What is missing

Oversight only works if humans can disagree. Rājadharma binds the one who decides.

आपद्धर्म *āpad-dharma*

Recognising the exceptional case, where ordinary rules no longer hold. *Mahābhārata, Śāntiparvan*

Uncertainty / abstention

Selective prediction and learning to defer let the system abstain, or hand a case to a human, when it is likely to be wrong.

Madras et al. 2018; Mozannar & Sontag 2020; Geifman & El-Yaniv 2017

What is missing

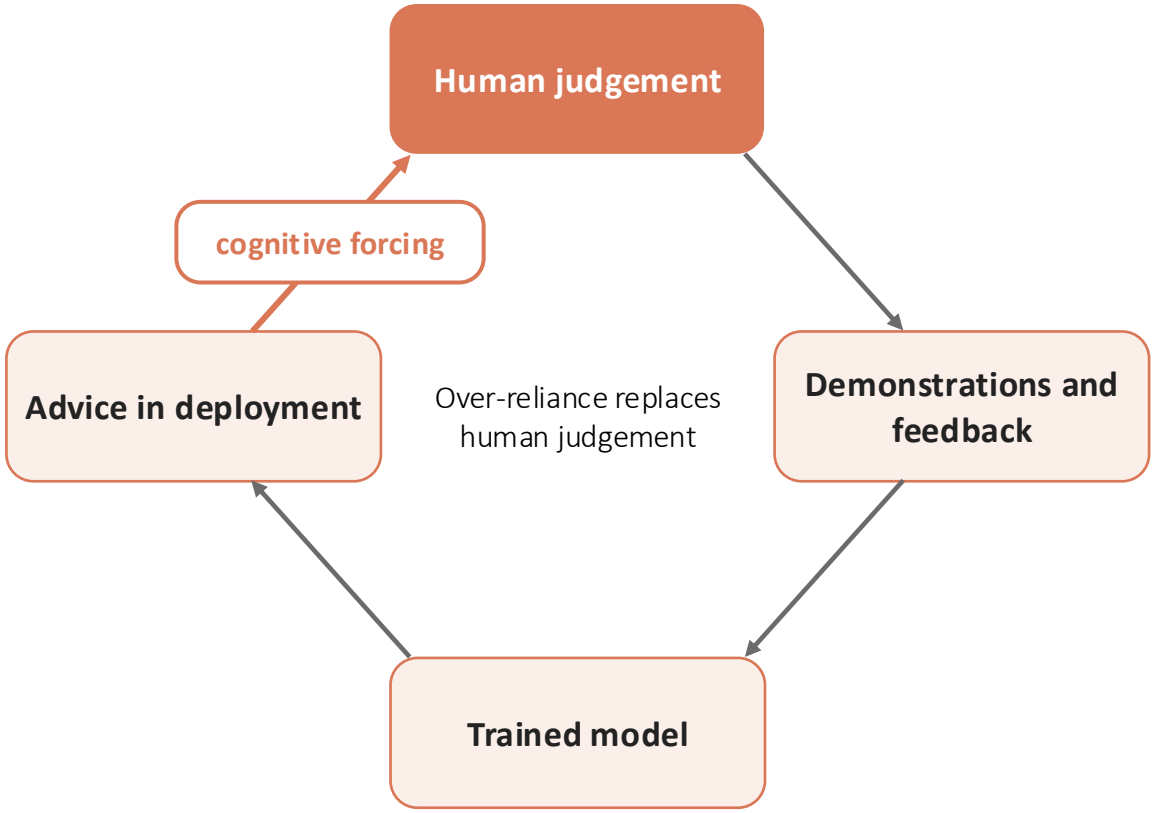
Abstention is triggered by predictive uncertainty. The question is whether this decision is the system's to make, not how confident it is.

Texts and papers cited in full on the final two slides

Governance: retaining human agency

श्रेयान्स्वधर्मो विगुणः परधर्मात्स्वनुष्ठितात्

Better one's own dharma, imperfectly performed, than another's performed well. Bhagavad Gītā 3.35 (Mahābhārata 6.25.35)



स्वधर्म svadharma: human agency

Overreliance. People over-rely on automated advice, and skills erode with disuse. Requiring a judgement before the AI's recommendation reduces over-reliance.

Croskerry 2003; Bainbridge 1983; Parasuraman & Riley 1997; Buçinca et al. 2021

What is required. Design systems so that humans must actively engage with the decision rather than simply accept the model's recommendation.



Image source: https://vediccosmos.com/yaksaprasnah-continued/

BACK AT THE LAKE

The Pandavas drank before they answered.

AI can be aligned to objectives, principles, and preferences. Dharma requires us to discern which should govern here.

1. **Whose path?**
   Make the provenance of demonstrations explicit.
2. **What is hidden?**
   Learn the reasons for decisions, not only the behaviour.
3. **When to step back?**
   Defer by design, and keep human judgement active.

महाजनो येन गतः स पन्थाः

The path the great have walked; that is the path to follow. Mahābhārata, Vana Parva 313.117