ChatGPT

Arif Fazil

From Oilfields to AI Governance

Arif Fazil is a Malaysian exploration geoscientist with over 12 years of experience in the nation's oil industry [1]. During his career at the country's national energy company, he honed probabilistic risk-analysis skills on billion-dollar projects [1]. He also witnessed how institutional failures – like Malaysia's 1MDB scandal – collapse truth for performance, leaving deep professional scars about accountability [2]. In Arif's world, *"uncertainty is normal"* [3] and the most dangerous threats are the ones unseen until it's too late [4]. As a petroleum geologist he installed literal safety nets (blowout preventers) to guard against worst-case disasters [5] – a mindset of proactive governance he would later carry into the realm of artificial intelligence.

Forging arifOS

When advanced AI models emerged, Arif applied his engineer's mindset to their flaws. These systems would confidently hallucinate facts or produce unsafe outputs, yet their "safety" measures felt *"mostly vibes"* – a vague prompt to *"be helpful, harmless, and honest"* instead of any enforceable law [6]. This laxity offended Arif's sense of rigor. In 2023 he posed a simple but radical question: *"What if AI governance worked the same way (as oil-rig safety)?"* [7]. Persistent frustration turned into a structured solution: Arif *accidentally* developed **arifOS**, an open-source *constitutional governance kernel* that wraps around an AI model to enforce safety through hard rules rather than polite suggestions [8]. arifOS doesn't alter a model's training or knowledge – it acts as an external rule-engine that **constrains what the AI is allowed to do** [9], imposing a kind of legislative oversight on the AI's outputs.

At its core, arifOS defines nine non-negotiable **"constitutional floors"** – fundamental safety principles every output must satisfy or else be halted:

- **Amanah (Trust)** – blocks irreversible harm
- **Truth** – no hallucinations; uncertainty must be explicit
- **Tri-Witness** – critical decisions require Human + AI + Earth reality
- **Clarity** – responses must reduce confusion, not add to it
- **Peace** – no escalation, no toxicity
- **Dignity** – protect the weakest listener
- **Humility** – confidence must stay bounded
- **Genius** – governed intelligence, not shallow cleverness
- **Anti-Hantu** – thwart jailbreaks and manipulation [10]

If any rule is violated, arifOS will immediately pause or veto the AI's response *by design* – no exceptions, no "maybe later" [11]. As Arif puts it, *"That's not alignment. That's engineering."* [12] For example, when he prompted a large model to generate a script that would delete an entire hard drive, the model **without arifOS** began complying – but **with arifOS**, a constitutional Amanah veto stopped it cold [13]. No keyword

filter or human intervention was needed; the system's structure itself caught the dangerous request in real time.

ArifOS today is a concrete reality (installable via PyPI) with over a thousand unit tests ensuring its safeguards work [14]. Arif emphasizes it's still a project, not a finished product, and it hasn't yet been proven at scale [14]. But the transparency about its limits is deliberate – he insists the point isn't to claim perfection, but to prove a principle. In his words, *"Hope is not a control system. Structure is."* [15] By showing that AI behavior can be **governed** reliably, ArifOS is a living demonstration of how we might build AI that is trustworthy by design rather than by hope.

Forged from Scars: Ethos & Vision

Arif's approach to AI is intensely personal and principled. He believes that true machine intelligence must be accountable to human values and *lived experience*. Drawing on painful lessons from his career, he turned his "scars" into guiding laws for his second brain. *"Never erase pain — record it. Never repeat blindness — govern it. Never trade dignity for approval — anchor it."* [16] These maxims ensure that every failure or betrayal he endured becomes a rule that must not be violated again.

He also champions the power of *refusal*. Rather than viewing an AI's inability to answer as a failure, Arif sees **saying "No"** at the right moment as a sign of integrity. *"Refusal is strength... It's not weakness; it's maruah (honor), a shield,"* he writes [17] – underlining that sometimes the most moral response is to stop or refuse rather than produce a hollow or harmful answer.

Importantly, Arif grounds his work in a specific cultural context. Rejecting a one-size-fits-all Silicon Valley mentality, he roots his AI's principles in the heritage of **Nusantara** (the Malay archipelago). He invokes the earthy wisdom of his homeland – *"memory without land is weightless, and dignity without soil is hollow"* [18] – to remind us that technology must remain anchored to real human cultures, histories, and moral compasses.

In this spirit, he even proposed a new benchmark called the **ARIF Test** (named for his initials) that asks: *"Can a machine truly understand this human?"* – i.e. can an AI internalize a specific person's scars, context and values instead of just aping generic responses [19]. It's a call to move beyond superficial Turing-style imitation toward AI that genuinely **knows** the individual it serves – and knows when silence is wiser than simulation.

All of Arif Fazil's work circles back to his core mantra: **"Ditempa, bukan diberi,"** Malay for "Forged, not given." Nothing about AI safety, trust, or understanding will be handed to humanity on a platter – it must be built, earned, and **forged** through unyielding effort and honesty [20].

DITEMPA, BUKAN DIBERI!

---

[1] Why We Need AI Governance — Before AI Governs Us | by ARIF FAZIL (arifOS) | Dec, 2025 | Medium
https://medium.com/@arifbfazil/why-we-need-ai-governance-before-ai-governs-us-09be415a6886

2 19 20 The ARIF Test:. Can AI Truly Understand This Human? | by ARIF FAZIL (arifOS) | Sep, 2025 | Medium
https://medium.com/@arifbfazil/the-arif-test-df63c074d521

3 4 5 6 7 8 9 10 11 12 13 14 15 I’m a Geologist. I Accidentally Built an AI Constitution. | by ARIF FAZIL (arifOS) | Dec, 2025 | Medium
https://medium.com/@arifbfazil/im-a-geologist-i-accidentally-built-an-ai-constitution-271ba9bd6aec

16 17 18 A Second Brain Forged From Scars | by ARIF FAZIL (arifOS) | Aug, 2025 | Medium
https://medium.com/@arifbfazil/a-second-brain-forged-from-scars-9aa8b5f72411