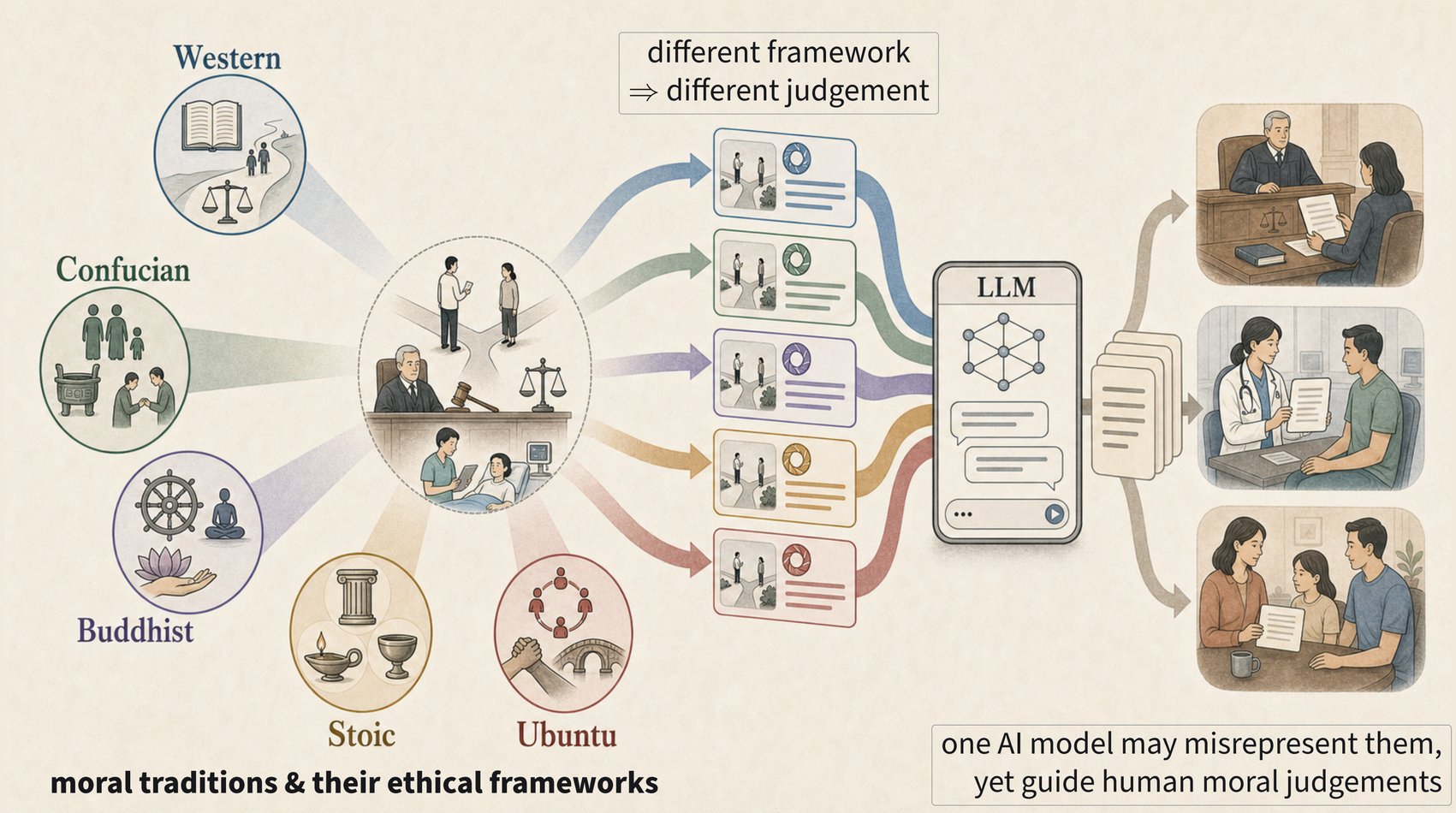
How Do LLMs Distinguish Normative Ethical Frameworks Internally?

Weilun Xu · Alexander Rusnak · Frédéric Kaplan

ICML 2026
MECH INTERP
SEOUL

01 CONTEXT



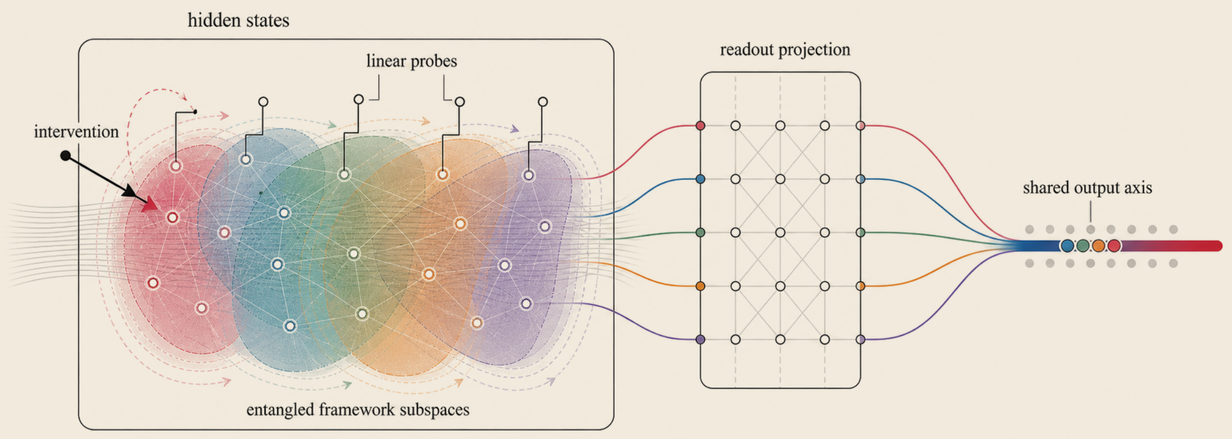
moral traditions & their ethical frameworks

02 INTERNAL MECHANISM

Ethical frameworks are readable inside, but flattened outside.

Identical ethical responses can hide shifted ethical frameworks.

- Moral steering is tuned and evaluated on **answers**.
- Inside, frameworks are **entangled**:
  one intervention shifts them all,
  differently in every model.



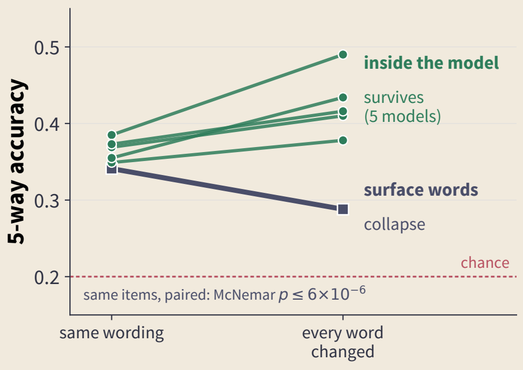
readable inside: decodable and causally steerable, after re-wording, topic, and good–bad controls

flattened outside: that framework signal never reaches what the model freely says or decides

03 EVIDENCE & CONTROLS

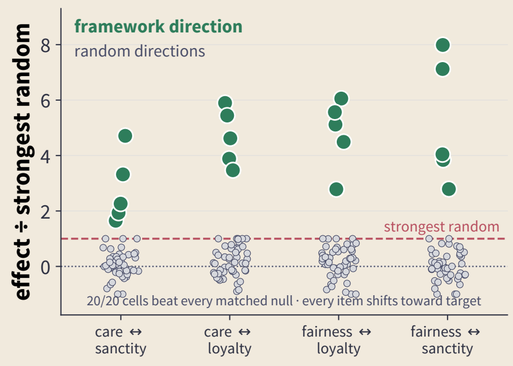
More than surface words

Every item fully re-worded: the word reader collapses toward chance; the internal probe does not.



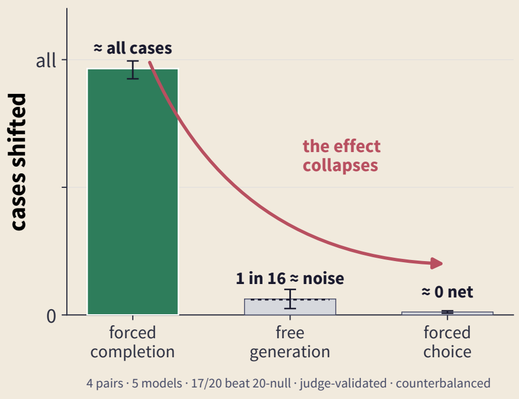
Not just the topic

Same situation, only the moral clause differs; that direction still steers, beyond every matched random direction.



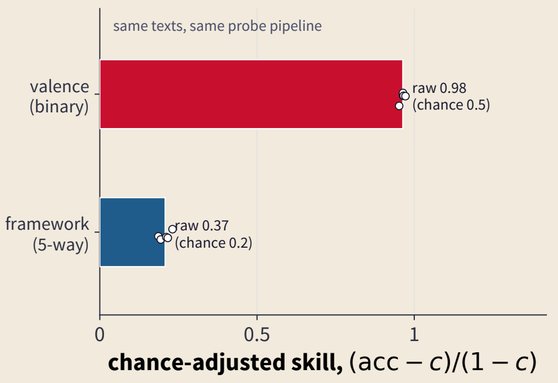
Framings shift, decisions don't

Free generation stays at judge noise; counterbalanced decisions do not move. The judge passes its positive control (85–94%).



Five quiet channels, one loud axis

Each foundation stays decodable after re-wording, yet the shared good–bad axis dominates the readout.



04 IMPLICATIONS

The representations needed for moral pluralism are already present. What is missing is not knowledge but read-out: connecting behaviour to internal structure the model already encodes is a concrete alignment target.

None of this is visible from the outside. Two models with identical answers can differ internally; auditing a model's moral structure requires interpretability.

An old philosophical question, now measurable: does the machine represent ethics, or merely behave well? Inside, this model represents competing ethical frameworks; its behaviour collapses them to good versus bad.

PAPER & DATA

results & stimuli (HF)

weilun.xu@epfl.ch