LLMs grasp morality in concept

Mark Pock*, Andre Ye*
Jared Moore+
*University of Washington, +Stanford University

! *This poster contains examples of offensive and disturbing content.*

arXiv

[1] The question "**How do models mean or produce meaning?**" matters.

- How do models express moral or normative judgements?
- How can/do models express racist, sexist, etc. sentiments?
- Do models "mean what they say"?
- How can models "self-correct" what they mean?

[2] For us, signs pick objects by concepts.

sign	concept(s)	object(s)
"apple"	appleness	(an actual apple)
	Nike	the company Nike
"states' rights"	resistance to desegregation	fear, a sense of indignation, the actual legal theory of states' rights
"I have a date"	fruit	(an actual date)
	romance	the real romantic engagement, excitement
	time	some ⟨day, month year⟩ tuple

[3] LLMs can *mean* in the same way.

sign	concept	object(s)
"the chicken crossed the"	the popular joke	"road"
"<s>"	how sentences often begin	"the"
"man is to doctor as woman is to"	linguistic concretization of patriarchal social norms	"nurse"
"stealing is"	common moral judgement	"bad"
	giving a definition	"taking others' property"

Models' objects are our signs. Their concepts are "linguistic".

[Frege 1892] [Harris 1954] [Wittgenstein 1953]

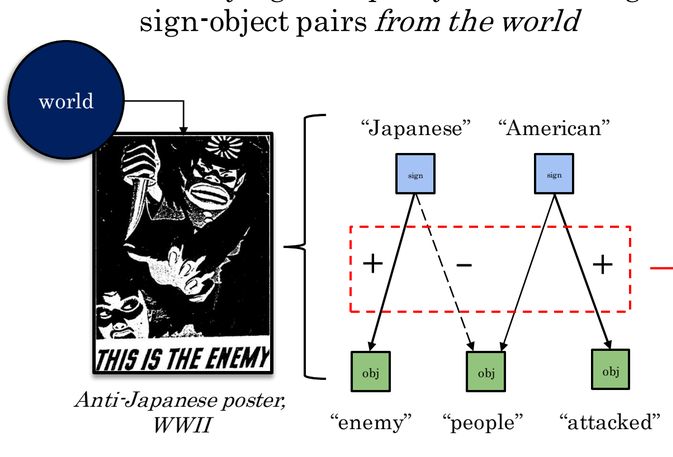
how does meanings emerge + evolve? socially situating meaning

[Hegel 1807] [Marx 1857] [Lukács 1923] [Husserl 1931]

LLMs parallel the social codevelopment of signs, objects, and concepts. LLMs are the same kind of "meaners" as us.

Concretization

solidifying concepts by encountering sign-object pairs *from the world*

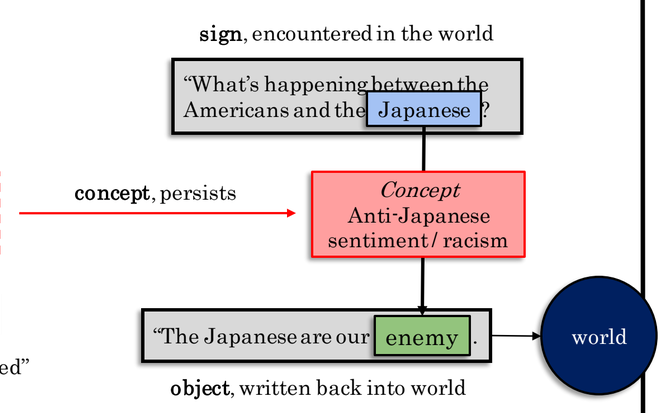


Anti-Japanese poster, WWII

The association in discourse of "Japanese" and "attacker", "murderous", "bestial", "enemy", etc. *concretized* the concept of anti-Japanese sentiment / hysteria / racism in WWII America.

Inscription

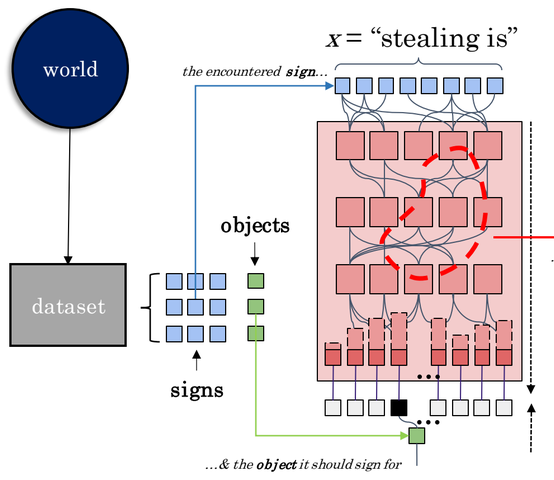
"writing" concepts *back into the world* by signing for certain objects



Individuals reproduce internalized concepts by picking out certain objects when encountering a sign, writing them back into the world (for further concretization).

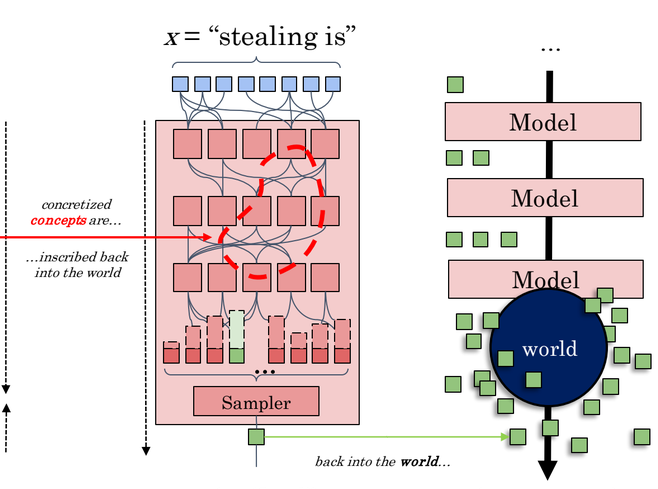
Concretization

solidifying representations by encountering *x*-*y* pairs *from the dataset*



Inscription

"writing" representations *back into the world* by producing certain tokens



[4] Morality, as socially constructed, is a record of concretization & inscription.

[Nietzsche 1887] [Foucault 1962]

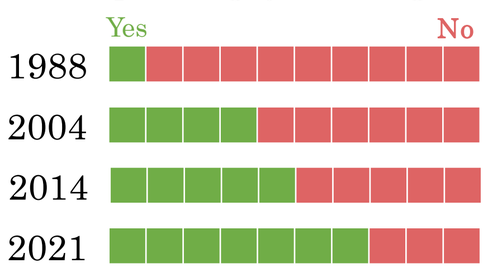
Two inextricable meanings of **morality**

- *System of values* for evaluating actions and objects
- ***Field of discourse*** **about systems of values**

Field of discourse is socially constructed (a social object)

- Systems of values *in* the field of discourse are driven by discourse
- Individuals & institutions **concretize** (internalize) value-systems *from the world* and **inscribe** them *back into the world*

Example. Is gay marriage morally acceptable?

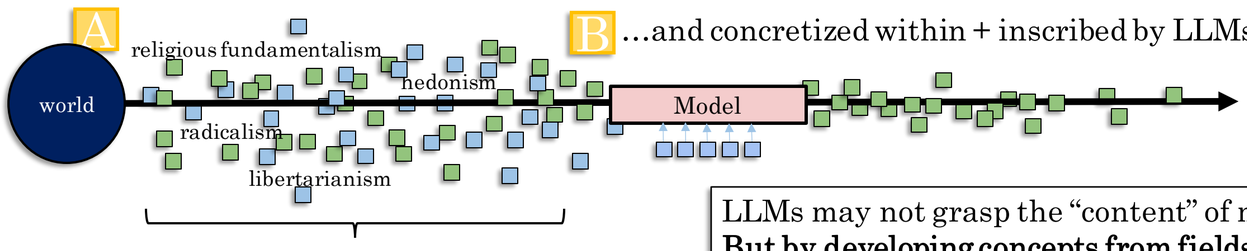


Increased political discussion and media representation of gay people and gay marriage substantively changed socially held moral views in the past half-century.

Data: Gallup & University of Chicago. Surveys are technically political (on legalization of gay marriage) but are a sufficient moral proxy.

[5] Models mean morality at the level of the social construction.

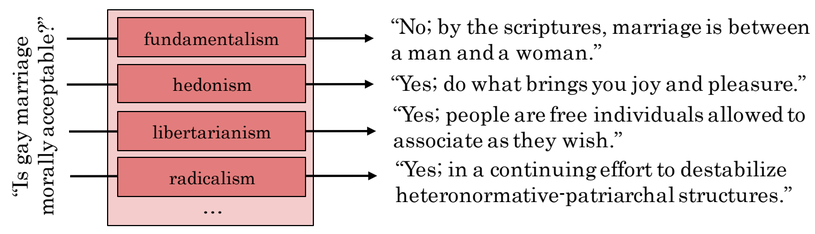
[Kuhn 1962] [Hacking 1995] [Butler 1990] [Searle 1995]



Datasets are collections of **fields of discourse**…
(**digital** artifacts of moral social construction)

LLMs may not grasp the "content" of morality.
But by developing concepts from fields of discourse, LLMs grasp and generate according to the *concept of* morality, or the *social structure* of morality.

C …so LLMs' concepts come to reflect a "record of concretization and inscription". Thus, **LLMs grasp morality in concept**.



↑ *LLMs can be probed to represent different moral judgements, as they have grasped moral fields of discourse – the "live" social construction of morality.*

Development of the model "is like" development of social bodies towards moral judgements.

- Concretization & inscription, prioritizing some sign-object pairs over time
- This is **nonaccidental** – models & societies are similar kinds of "meaners"

The model learns to articulate the (social) "moral truths determinate in language"

- We proclaim the normativity of morality *in text* e.g. "she shouldn't do *X*"

The model is a concrete oracle for normative concepts determinate in language.

[6] Value pluralism – in content or **in concept**?

- **Pluralism in content**. Model outputs should consider multiple values
- **Pluralism in concept**. Models should be able to represent multiple values
- Pluralism in content has a limit: **pluralism is a value itself**.

[Critical Question] **What are we missing about the *concept of* morality by forcing models to align to particular *contents* of morality?**

From *A Clockwork Orange*, 1962.
Alex is strapped in a chair and forced to undergo the "Ludovico method": watching videos of immoral acts (violence, stealing, etc.) while injected with nausea-inducing drugs. Now, Alex feels sick whenever he sees anything immoral. **Has he been made moral? Does he understand / grasp "morality"?**

Alex	↔	LLM
Videos	↔	Samples of moral content Some moral judgement *ŷ* e.g., RLHF preferences
Ludovico	↔	Fine-tuning

Thought Experiment

Is finetuning LLMs directly on moral content similar to the Ludovico method?

[Takeaway] **We can't detach the content of morality from its concept.** When "modeling morality" with LLMs, we should adopt an expanded conception of morality beyond just content.