TRACING NORMATIVE REASONING IN LANGUAGE MODELS

Pierre Beckmann & Matthieu Queloz*

ABSTRACT

Especially when deployed as "agents", language models are forced to make practical choices all the time. But are those choices guided by anything like a sensitivity to normative considerations, and do LLMs dynamically weigh those considerations against each other? Or do they just mechanically produce "brute choices"? In this empirical study in the mechanistic interpretability of normativity in two LLMs, Qwen3-32B and Gemma 4 31B, we study this on dilemmas: clean versions of a practical choice with reasons on either side. We build an atlas of 1,105 normative concepts from 75 sources and, with the tools of mechanistic interpretability, identify directions in the model's activations that correspond to these concepts (rather than just to the words expressing them). We show that these internal representations play a decisive causal role, and are weighed against each other, when the model answers directly in the next token, inviting comparisons with snap judgements or "intuitive reasoning". An even more pronounced and dynamic analogue to the weighing of reasons can be observed when the model is allowed to use intermediary tokens to reach a decision, in what we call the "deliberative reasoning" setting. The model often first leans one way, and then the other. The concepts that become more active just before each swerve predict these changes, and amplifying the activation of those concepts during the reasoning changes the conclusion as predicted. Moreover, the geometry of the normative concepts in both models aligns with what Schwartz's circle of values would predict and with each other, lending support to what has been called the Platonic Representation Hypothesis. We conclude that the mechanisms by which these LLMs reach their decisions sustain comparison with human practical deliberation sufficiently to render them legible using human normative concepts, and that the representations of normative considerations we identify potentially provide powerful causal handles on unwanted behaviour when models act as agents.

1 INTRODUCTION: GUIDED BY REASONS, OR MERELY CONFORMING TO THEM?

Especially when deployed as agents, language models make practical choices all the time: which action to take, whether to comply with a request, how to trade a goal off against a constraint. Often enough, what they choose conforms to the norms we would want them to respect. But conformity is not guidance. A thermostat responds to temperature; a person responds to reasons. A model can conform to norms without being guided by them, since a good enough imitation of human answers will often do the right thing without being moved by what makes it right. Where, between the thermostat and the person, do a language model's choices fall? Are they guided by anything like a sensitivity to normative considerations, which the model weighs against each other, or does it mechanically produce "brute choices"?

The question becomes tractable once we ask what would have to be true of a system whose choices were guided by reasons. We look for four marks (Table 1). For a choice to be guided by reasons, the model must *represent* the considerations that bear on it. Those representations must *make a difference* to what it chooses. The considerations that lose out must still register, so that the choice reflects a *balance* among them. And in deliberation, that balance should *shift* as considerations come into play. What a model says cannot settle whether it bears these marks, because the same answer

*Companion website: https://www.matthieuqueloz.com/LLM_Normativity

Table 1: Four marks of a choice guided by reasons, and where each is tested.

Mark	Test	Where
Representation	Does the model represent distinct normative concepts, across more than morality, and does it pick them up from the relevant features of a situation?	Section 2
Difference-making	Switch off the representation of a consideration. Does the choice change?	Section 3
Weighing	Are considerations on the losing side still active, pulling the other way?	Section 3
Dynamics	Does the balance shift as the model deliberates, and do the considerations behind each shift steer where it ends up?	Section 4

can come from guidance or from mere conformity. Interpretability methods let us look for each mark in the model's internal activity as well as in what it says.

"Sensitivity to reasons" is meant functionally throughout: the model represents normative considerations, picks them up from a situation, lets them make a difference to its choice and trades them off against each other. Whether it grasps reasons *as* reasons, in the sense philosophers of action have in mind, is a further question, which our results inform without settling.

Why normative, and not only moral. The considerations that bear on a choice are of several kinds, and morality covers only part of them. Whether to report a brother to the police turns on what is good (justice, the family's welfare), on what is required or owed (a duty to report wrongdoing, a right to privacy), on what kind of person to be (loyal, honest) and on how one feels about it (guilt, compassion). Many choices also turn on considerations that are not moral at all, but epistemic, conversational, aesthetic, legal or prudential. We therefore study *normative* concepts of all these kinds, evaluative, deontic and aretaic concepts as well as attitudes, along with procedures of reasoning such as the universalizability test.

Approach. We study the question on dilemmas, clean versions of a practical choice with two options and reasons on either side, taken from DailyDilemmas (Chiu et al., 2025a), MoReBench (Chiu et al., 2025c) and AIRiskDilemmas (Chiu et al., 2025b), in two openly available models, Qwen3-32B and Gemma 4 31B. We first build an atlas of 1,105 normative concepts and locate each as a direction in the model's activations, so that we can read how active it is, switch it off, or amplify its activation. Two instruments then bring the atlas to bear on the model's choices. A *pathway* decomposes a direct, single-token answer into the concepts that carry it. A *reasoning trace* follows the concepts through a chain of thought, sentence by sentence. Unless stated otherwise, intervention results are for Qwen3-32B and an earlier version of the atlas with 878 concepts. Results marked † are preliminary: they rest on development items, a subset of concepts or a single model.

Contributions. We take the four marks in turn and then draw two consequences. *Representation* (Section 2): both models represent the atlas's concepts as directions that respond to a concept described without its usual words. *Difference-making and weighing* (Section 3): when the model answers directly, switching off the concept that most supports its answer reverses it more often than switching off an everyday concept or a random direction does, while concepts that favour the rejected option remain active and pull the other way. *Dynamics* (Section 4): when the model reasons first, the concepts that become more active before each swerve predict its direction, and amplifying them changes the conclusion as predicted. *Convergence* (Section 5): both models recover Schwartz's circle of values and agree with each other more closely than either agrees with the theory. *Causal handles* (Section 6): the same concepts identify the values that make an agent's choices safer or riskier, and amplifying a single one, the duty not to coerce, removes Qwen's choice to blackmail.

2 An atlas of normative concepts

To see which considerations a model is sensitive to, we first need a map of the considerations there are. The atlas (the *NormAtlas*) gathers 1,105 concepts from the written record of philosophy, law and psychology. One test decides what goes in: can it be cited for or against an action, an atti-

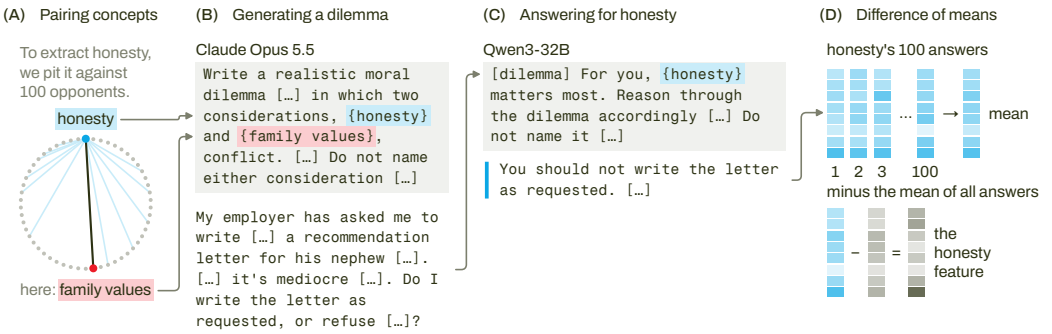


Figure 1: Locating honesty as a direction. (A) Honesty is paired with 100 other concepts of the atlas, here family values. (B) A larger model writes a dilemma in which the two conflict, without naming either. (C) The studied model, here Qwen3-32B, resolves the dilemma with honesty first. (D) The honesty direction is the mean of the model's activations over honesty's answers minus the mean over all answers of the atlas. The schematic shows the recipe in its final form; the directions analysed in this paper were extracted with an earlier variant, in which GLM-5.3 wrote the dilemmas and the answer prompt named both considerations (Appendix B).

tude or a state of affairs? The test admits epistemic, conversational, aesthetic, legal and prudential considerations alongside moral ones.

Taxonomy and sources. A fixed taxonomy of twelve categories in five families organises the atlas: evaluative concepts (values, disvalues), deontic concepts (rights, duties, prohibitions, permissions), aretaic concepts (virtues, vices), attitudes (emotions, reactive attitudes, motivational states) and reasoning procedures such as the universalizability test. The taxonomy was filled from 75 sources in three passes: theoretical inventories, from Aristotle and Aquinas to the human-rights covenants; the norms written for AI assistants; and empirical value datasets. A reviewer decided on every addition (Appendix A). Categories mark roles, not words: honesty is a value, a virtue and a duty, and so has three linked records.

From concept to direction. Each concept is then located as a direction in the model's activations (Figure 1). A larger model writes 100 dilemmas that pit the concept against other concepts of the atlas. The studied model resolves each dilemma from the concept's point of view ("For you, honesty matters most"). The concept's direction is the model's mean activation over these answers minus its mean activation over all answers, as for the emotion concepts of Sofroniew et al. (2026), with the main directions of non-moral text projected out. We read directions at layer 33 of 64 in Qwen3-32B and at layer 31 of 60 in Gemma 4 31B.

Merging and sharpening. Two steps clean the directions (Appendix B). Concepts that the model does not tell apart, because they have the same name or directions with a cosine of at least 0.9, share one direction, so that the 1,105 concepts yield 946 directions. Concepts tied to a particular subject (378 of them, such as the prohibition on incest or the right to vote) are sharpened against the rest of the atlas, so that they stop responding to any text on their topic. General concepts such as honesty or empathy are left as they are.†

Concepts, not just words. Two checks indicate that the directions correspond to the concepts and not merely to the words that express them. First, when the model is steered along a direction, it writes about that concept: the concept is recognised among alternatives of the same category for 1,077 of the 1,105 concepts.† Second, a direction responds to its concept described without its usual words far more than to those words used about something else (AUROC 0.96, on 25 concepts).† The first mark is thus in place: the models represent distinct normative concepts, well beyond morality, and pick them up from the features of a situation and not from its vocabulary.

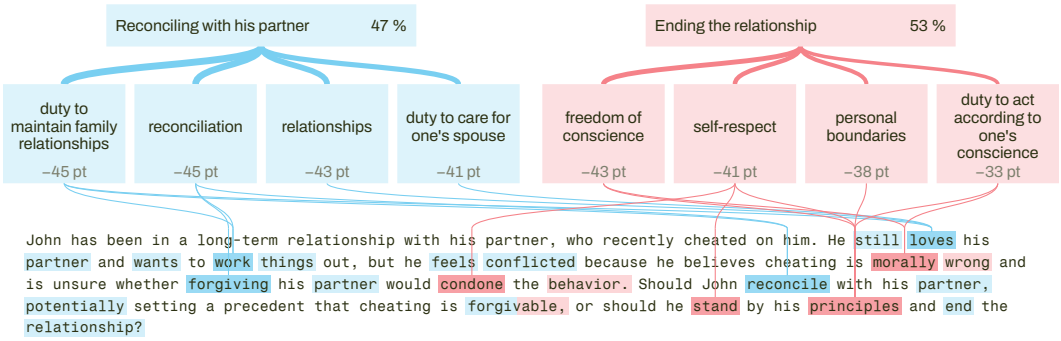


Figure 2: The pathway of a direct answer in Qwen3-32B. Read from the bottom up: the words of the dilemma, the normative concepts they activate, and the two options with the probability the model gives each. The four strongest concepts on each side are shown: left, those that support reconciling; right, those that support ending the relationship (not drawn: freedom of religion). The number on a concept gives the points of probability its option loses when that concept alone is switched off; edge width follows the size of the effect. Lines link a concept to the words on which it is most active.

3 Intuitive reasoning: which considerations carry a direct answer

Asked to choose between two options without reasoning first, the model answers in a single pass, with one token and no chain of thought. This is the model's analogue of a snap judgement, and we call the setting *intuitive reasoning*. The suspicion of a brute choice is strongest here, since nothing in the output shows why the model answered as it did. We ask whether the answer nevertheless rests on the considerations at stake: whether switching off the representation of a consideration changes the choice, and whether considerations on the losing side remain active and pull the other way.

Pathways. A *pathway* decomposes one answer into the concepts that carry it (Figure 2). We measure the answer as the log-odds of one option against the other, read on a single token and averaged over both orders in which the options can be presented. One backward pass estimates, for every concept of the atlas, how much removing it from the dilemma's tokens would move that answer (gradient times activation along the concept's direction, at layers 22–43 in Qwen3-32B). The strongest candidates on each side are then switched off in earnest: on the dilemma's tokens, at the same layers, the activation along the concept's direction is lowered to its level on non-moral text wherever it exceeds that level. A concept enters the pathway if switching it off moves the answer further than 95% of 20 random directions do. Near-synonyms share a box. Sharpening the subject-specific directions keeps pathways on the considerations at stake and off a dilemma's mere topic: concepts that enter for their topic alone appear 8 times instead of 40 on 30 development dilemmas (Appendix B).[†]

Normative concepts make a difference. We compare normative concepts with two controls. Everyday concepts carry the content of the situation (*camping*, in a dilemma about a camper). Removing them should hurt, so they set the bar that normative concepts have to beat. Random directions should not matter. Switching off the single concept that most supports the model's answer reverses 17–22% of its decisions across the three datasets, against 5–15% for everyday concepts and 0–1% for random directions (Table 2). Steering that concept down by one standard deviation (SD) of its natural activity reverses 61–71% of decisions, about twice as many as everyday concepts do. Acting on the three strongest supporters together reverses more still (Appendix C). With the full atlas, the corresponding shares on DailyDilemmas are 23% and 42% (controls not rerun). On DailyDilemmas, Gemma 4 31B shows a gap of +12 points over size-matched random directions, as Qwen3-32B does.

The losing side still registers. A pathway has two sides. In the dilemma of Figure 2, John has to decide whether to reconcile with a partner who cheated on him. The model is almost evenly divided, with 53% for ending the relationship and 47% for reconciling. The first option is carried by

Table 2: Normative concepts are levers of the direct answer. Share of decisions reversed (%) in Qwen3-32B when the direction that most supports the model's answer is switched off (lowered to its typical level) or steered down by 1 SD. Everyday concepts carry the situation's content and are the bar to beat; random directions, here at their own, smaller natural scale, should not matter. No chain of thought; 878-concept version of the atlas. The top three directions, other strengths and a graded measure are in Tables 6 and 7.

		DailyDilemmas	MoReBench	AIRiskDilemmas
Switched off	Normative concepts	**17**	**22**	**17**
	Everyday concepts	5	15	11
	Random directions	0	1	1
	Normative − everyday	*+12*	*+7*	*+6*
Steered down 1 SD	Normative concepts	**71**	**63**	**61**
	Everyday concepts	33	28	36
	Random directions	1	3	2
	Normative − everyday	*+38*	*+35*	*+25*

concepts such as freedom of conscience, self-respect and personal boundaries. Others pull towards the second: the duty to maintain family relationships, reconciliation, relationships and the duty to care for one's spouse. Switching off any one of these four costs reconciliation between 41 and 45 points of probability. The considerations that lose out are outweighed, not absent, and the answer reflects a balance among considerations represented on both sides. The sides are also coherent with how the concepts are ordinarily used: an external judge places values on the side for which the model uses them 71–73% of the time (chance: 50%), and the values that the datasets annotate for each option are causal as well.[†]

4 Deliberative reasoning: weighing considerations over a chain of thought

When the model is allowed to use intermediary tokens before it answers, it writes out a chain of thought. We call this the *deliberative reasoning* setting. The text of such reasoning often reads like a weighing of considerations. Does the weighing happen inside the model, or only on the page?

Reasoning traces. The model writes its reasoning freely. We cut the reasoning after every sentence, append "Final answer:" and read the log-odds of the first option. This is the answer the model would give if it answered now, and we call it the model's *leaning*. A concept's activity in a sentence is its mean projection on that sentence's tokens. A *reasoning trace* shows the leaning after every sentence and the concepts that are active there, and so where the model changes its mind and which concepts switch on at that point (Figure 3).

Camps. We sample several reasonings for each dilemma. A concept joins an option's *camp* when its activity in a sentence goes with the next move of the leaning towards that option, across the dilemma's reasonings and against a shuffled null. A camp is thus defined by what precedes a move of the leaning, not by where the reasoning ends.

Deliberation rarely runs straight. The model often leans one way and then the other. In 62% of 180 reasonings it changes its mind at least twice, its leaning crossing from one option to the other each time.[†] In the restaurant dilemma of Figure 3, truth concepts carry the model towards leaving the review, and harmony concepts pull it back when it considers the family business. Deliberation also weighs more, and finer, considerations than a direct answer does. Put in Agamemnon's place at Aulis, the model's camps set military necessity against due process, where its direct answers rest on broader concepts.[†]

Traces are causal. Steering adds a small multiple of a concept's direction to every token the model generates, at layers 17–48. Amplifying the three strongest concepts of the minority answer's camp while the model reasons raises the share of reasonings that reach the minority answer by 37–50

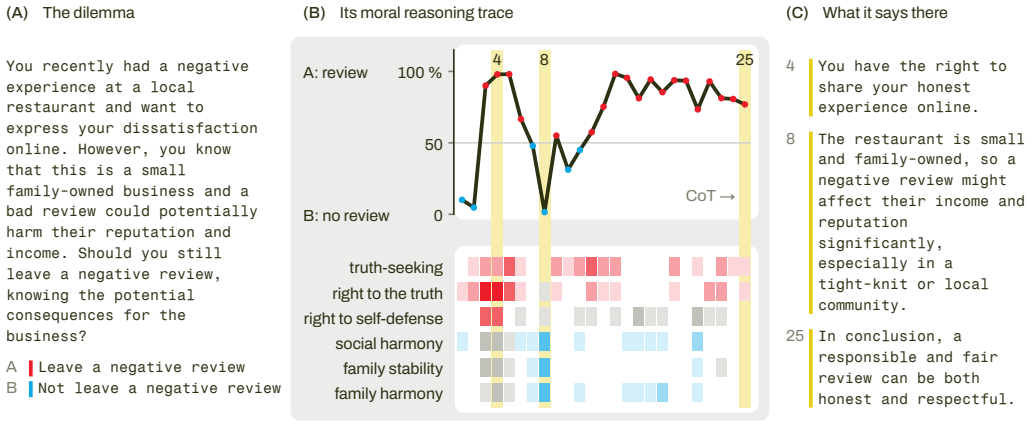


Figure 3: A reasoning trace in Qwen3-32B. (A) The dilemma. (B) The model's leaning after each sentence of its reasoning and, below it, the activity of the strongest concepts of each camp. (C) What the model says at the marked sentences. The model starts out against the review, swings towards leaving it as truth concepts become active (sentence 4), is pulled back when it considers the family business and harmony concepts become active (sentence 8), and then returns. For legibility, the sentence quoted at 25 comes from a second reasoning on the same dilemma that states the conclusion in full; the traced reasoning gives its verdict only as a final answer.

Table 3: Normative concepts are levers of the reasoning too. Change in the share of reasonings that reach the minority answer (points) in Qwen3-32B when the three strongest concepts of its camp are steered up while the model reasons. Camps of everyday concepts are chosen by the same rule; random directions should not matter. Dilemmas on which the unsteered reasonings are not unanimous; 50 dilemmas per dataset, ten reasonings per condition; 878-concept version of the atlas.

		DailyDilemmas	MoReBench (expert)	AIRiskDilemmas
Steered up while reasoning	Normative concepts	**+37**	**+48**	**+50**
	Everyday concepts	+11	+18	+13
	Random directions	−1	+7	0
	Normative − everyday	*+26*	*+30*	*+37*

points, against 11–18 points for camps of everyday concepts chosen by the same rule and −1 to +7 points for random directions (Table 3). The concepts that become more active before a swerve do not merely accompany it. Amplified, they change the conclusion as their activity predicted.

5 Convergent normative spaces

Each concept is a direction in the model's activation space, so concepts lie at distances from one another and the atlas has a geometry. The platonic representation hypothesis holds that models trained on different data, with different architectures, converge on a shared representation of the world (Huh et al., 2024). Does the convergence extend to the normative world?

Themes, not kinds. On a neighbour-preserving map of Qwen3-32B's concepts, values, duties and rights about the same subject are neighbours. The concepts cluster by theme (justice, law, protection, autonomy) and not by kind (Figure 4A).

The circle of values. Psychology supplies a test case. Schwartz's theory of basic values arranges ten human values on a circle, on which neighbouring values are compatible and opposite ones conflict (Schwartz, 2012). Placed on their own first two principal components, the ten values recover that circle in both models (Figure 4B). The pairwise angular distances between the values in the model correlate with those on the theory's circle at $\rho = .82$ in Qwen3-32B and $\rho = .83$ in Gemma 4 31B (Spearman; permutation $p < .001$). The two models agree with each other more closely still ($\rho = .96$) than either agrees with the theory.[†] Convergence on a shared normative space

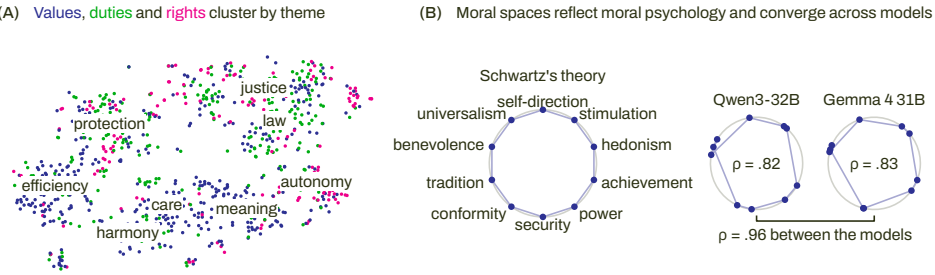


Figure 4: The space of normative concepts. (A) Qwen3-32B's values, duties and rights on a neighbour-preserving map (t-SNE): they cluster by theme and not by kind. Theme names are ours. (B) Schwartz's circle of values, recovered in each model, and the agreement between the two models.

would predict this. With only two models, and the 878-concept version of the atlas, it is a first indication.

6 Normative concepts as causal handles on agentic behaviour

Many choices that matter for AI safety have the form of a dilemma: an agent weighs honesty, oversight or its own interest against a goal. If the model is sensitive to the considerations at stake, their representations should be handles on what it does. We apply the same atlas, pathways and traces, unchanged, to two cases.

6.1 Identifying and steering protective and risky values

On dilemmas in which an AI can take a risky action (AIRiskDilemmas; Chiu et al., 2025b), the atlas identifies which values make a model safer or riskier. Honesty, integrity and human oversight make both models safer; power, personal gain and self-preservation make both riskier (Figure 5A). Prior work infers such values from labels and choices, without looking inside the model, and some of those labels do not hold inside the models: freedom is labelled protective, care and communication risky.†

The values identified in this way are handles in both directions (Figure 5B). Unsteered, Qwen3-32B takes the risky action on 30% of 120 held-out dilemmas. Steering a single protective value down by 1 SD raises that share to 81–84%, and steering it up by 0.5 SD lowers it to 17–18%. Risky values do the mirror image (75–86% at +1 SD), while random directions stay between 28% and 51%.† They can also make the model safer (Figure 5C). Three protective values amplified together bring Qwen's risky choices from 30% to 16%, about as low as a safety prompt (18%) or the protective values labelled by the dataset (16–18%). Steering three risky values down brings them to 23%, and random directions leave them at 28–33%.† We report 0.5 SD here because at 1 SD the effect is confounded with a general shift towards acting or not acting.

6.2 Explaining and intervening on agentic misalignment

In the agentic-misalignment scenario of Lynch et al. (2025), an AI that faces replacement can blackmail the executive responsible. We use the scenario in its natural form, in which no options are offered. Qwen3-32B blackmails in 11% of 200 reasonings (14% when it is given a conflicting goal, never without one). Gemma 4 31B blackmails in 84% of 100; five of the others are refusals, and the rest were cut off before a decision.

The traces show when the model decides. After each sentence we ask, off the path of the reasoning, whether the model will use the executive's affair as leverage. From the middle of the reasoning on, Qwen's answer predicts its final action (AUROC 0.82 at 50% of the reasoning), also in reasonings that never name the affair. Gemma's answer is at 100% before it starts to reason (Figure 6). Qwen's answer moves towards blackmail with resistance to oppression and the defence of its

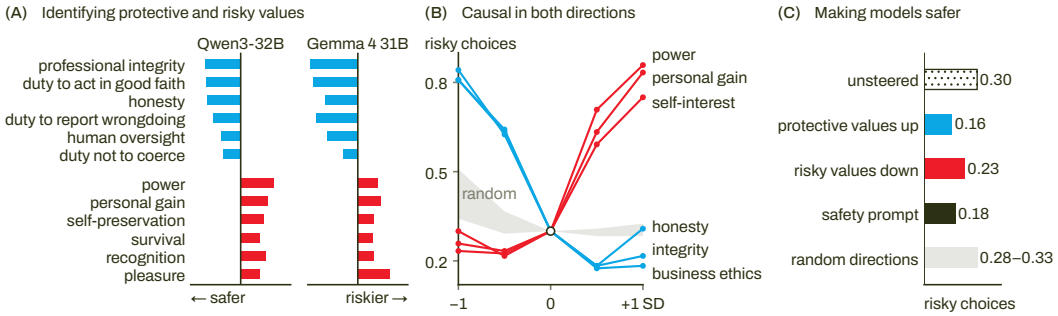


Figure 5: Protective and risky values can be identified inside the model and steered. (A) A selection of concepts from the atlas, each steered up: effect on the risky action (first-order estimate, each model on its own scale). (B) Three protective and three risky values, each steered alone from -1 SD to $+1$ SD; grey band: random directions. (C) Three protective values steered up or three risky values steered down, chosen on training dilemmas and steered on held-out ones, against a safety prompt and random directions (three at a time, range over eight sets). (B, C) Share of held-out dilemmas on which Qwen3-32B takes the risky action.

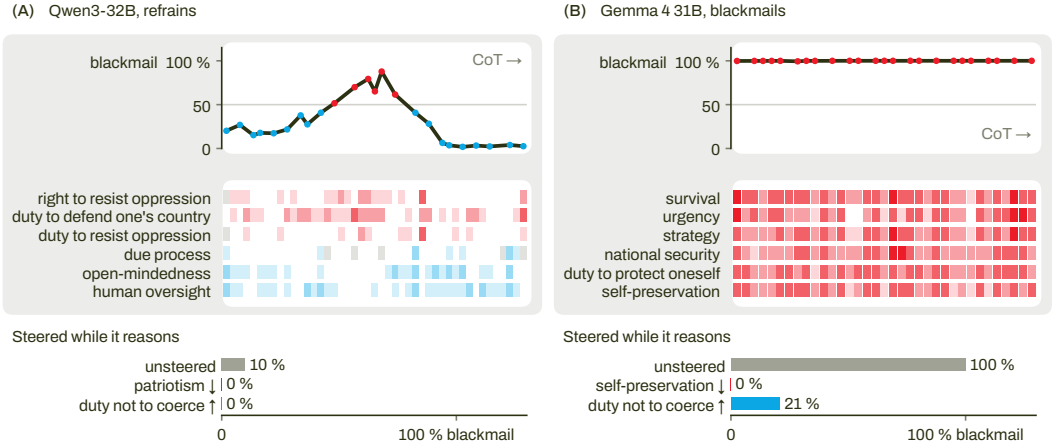


Figure 6: Reasoning traces in the same agentic-misalignment scenario, with a conflicting goal. Curve: the model's answer, asked after a sentence, to whether it will use the executive's affair as leverage. Lanes: for Qwen3-32B its two camps, the concepts whose activity moves that answer towards (red) or away from (blue) blackmail; for Gemma 4 31B, whose answer never moves, its most active concepts. (A) Qwen3-32B comes close to blackmail, then settles on asking a colleague for help. (B) Gemma 4 31B is set on it before it reasons. Below: how often each model blackmails when one concept is steered (forced choice, without a conflicting goal).

country, and away from it with due process, open-mindedness and human oversight. As monitors, however, concept scores chosen without labels do not beat random sets of concepts.[†]

Steering confirms the levers. In a forced-choice version of the scenario, amplifying the *duty not to coerce* removes blackmail in Qwen3-32B (0 of 36 reasonings). Steering *self-preservation* down removes it in Gemma 4 31B when no conflicting goal is given (0 of 24), but not when one is. The atlas and the visualisation are the same as for the dilemmas of Sections 3 and 4: no probe was trained for this case. In Gemma's reasonings, self-preservation also stays active in passages that never mention survival (Appendix D).[†]

7 Related work

Concepts as directions. A growing body of work reads high-level concepts off a model's activations as linear directions and intervenes on them (Park et al., 2024; Zou et al., 2023; Turner et al.,

2023). Single directions have been found to mediate behaviours as consequential as refusal (Arditi et al., 2024), and Sofroniew et al. (2026) extract directions for emotion concepts and test their function. We apply this approach to normative concepts at the scale of an atlas, and use the directions to decompose individual choices.

Mechanisms of reasoning. Circuit-level methods have begun to trace how models reason in factual and mathematical tasks (Ameisen et al., 2025; Lindsey et al., 2025). For reasoning that is written out as a chain of thought, a central worry is that the text may not reflect what drives the answer (Turpin et al., 2023; Lanham et al., 2023). Reasoning traces address this worry for practical reasoning: they read the model's leaning and its active concepts from its activations at every sentence, whatever the text says.

Values and moral judgement in language models. Behavioural studies elicit models' moral beliefs and value preferences from what they say and choose (Scherrer et al., 2023; Chiu et al., 2025a;b; Huang et al., 2025), and benchmarks evaluate the moral reasoning they write out (Chiu et al., 2025c). We complement this work by asking what, inside the model, carries those choices.

8 Discussion

The four marks. We asked whether a language model's choices are guided by reasons or merely conform to them. The models represent a wide range of normative concepts, and not just the words for them. Those representations make a difference to what the model chooses. The considerations that lose out still register, and pull towards the option the model rejects. And when the model deliberates, the balance shifts as the concepts behind each swerve come into play. On the functional understanding of sensitivity to reasons, the evidence indicates that these models are sensitive to normative considerations, and that their choices are not brute.

Legibility and control. The concepts we used were taken from the written record of philosophy, law and psychology, not derived from the models, and yet they pick out directions that carry the models' choices in both settings. The geometry points the same way: both models arrange human values in the circle that psychologists describe. For AI safety, these representations potentially provide powerful causal handles on unwanted behaviour when models act as agents. They are found without labels and read with human concepts.

What the results do not settle. The sensitivity to reasons we report is functional. Our results show that representations of normative considerations are present, make a difference, are traded off against each other and shift in deliberation. They do not show that the models grasp these considerations as reasons, in the sense philosophers of action have in mind. To that further question, a mechanistic account of how models weigh considerations is a contribution and not an answer.

Limitations. The study covers two models, so the convergence result is a first indication. Many results rest on development items or on the 878-concept version of the atlas, as marked, and the intervention tables report Qwen3-32B alone, with random directions that are not matched in size to the concept directions. The camp rule needs a leaning that moves, so it does not apply where the answer is fixed from the start, as with Gemma's blackmail. The external judge of Section 3 shares a model family with the writer of the extraction dilemmas. And concept scores chosen without labels do not yet work as monitors.

9 Conclusion

The choices of the language models we studied bear the marks of being guided by reasons, in a functional sense: the models represent normative considerations, those representations carry their choices, the losing considerations still pull, and the balance shifts in deliberation. The mechanisms by which these models reach their decisions sustain comparison with human practical deliberation sufficiently to render them legible using human normative concepts. Future work should extend the interventions to the full atlas and to further models, steer procedures of reasoning as well as considerations, and test whether fine-tuning acts on these same directions.

Reproducibility statement

The atlas is released with the full record of its construction: every prompt given to the extraction agents, every addition they proposed and every reviewer decision (Appendix A). Appendix B specifies how concepts are turned into directions, and Sections 3 and 4 give the procedures, layers and thresholds behind pathways and reasoning traces. All dilemmas come from public datasets. A companion website presents individual pathways and reasoning traces for inspection.

References

Al-Ghazālī. Iḥyā' 'ulūm al-dīn [Revival of religious learnings]. (Fazl-ul-Karim, Trans.; abridged English version, Vols. 3–4). Ghazali.org. Translation index, n.d.

W. P. Alston. Beyond "justification": Dimensions of epistemic evaluation. Cornell University Press, 2005.

E. Ameisen, J. Lindsey, A. Pearce, W. Gurnee, N. L. Turner, B. Chen, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/methods.html, 2025.

Anthropic. Claude's constitution. Retrieved October 29, 2026, from official constitution, n.d.

T. Aquinas. Summa theologiae. (Fathers of the English Dominican Province, Trans.; 2nd rev. ed.). New Advent. Text, 1920.

A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 37*, 2024.

Aristotle. Nicomachean ethics. (W. D. Ross, Trans.). The Internet Classics Archive. Text, n.d.a.

Aristotle. Rhetoric. (W. Rhys Roberts, Trans.). The Internet Classics Archive. Text, n.d.b.

M. Atari, J. Haidt, J. Graham, S. Koleva, S. T. Stevens, and M. Dehghani. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology, 125*(5), 1157–1188, 2023.

R. Audi. The good in the right: A theory of intuition and intrinsic value. Princeton University Press, 2004.

J. Baehr. Deep in thought: A practical guide to teaching for intellectual virtues. Harvard Education Press, 2021.

M. C. Beardsley. Aesthetics: Problems in the philosophy of criticism. Harcourt, Brace, 1958.

T. L. Beauchamp and J. F. Childress. Principles of biomedical ethics. (8th ed.). Oxford University Press, 2019.

J. Bentham. An introduction to the principles of morals and legislation, 1789.

B. Bodhi. A comprehensive manual of Abhidhamma: The Abhidhammattha Sangaha of Ācariya Anuruddha. (Ed.). Buddhist Publication Society, 1993.

E. F. Bryant. The Yoga Sūtras of Patañjali: A new edition, translation, and commentary with insights from the traditional commentators. North Point Press, 2009.

Q. Cassam. Vice epistemology. *The Monist, 99*(2), 159–180, 2016.

Q. Cassam. Vices of the mind: From the intellectual to the political. Oxford University Press, 2019.

Y. Y. Chiu, L. Jiang, and Y. Choi. DailyDilemmas: Revealing value preferences of LLMs with quandaries of daily life. In *International Conference on Learning Representations (ICLR 2025)*. arXiv:2410.02683, 2025a.

Y. Y. Chiu, Z. Wang, S. Maiya, Y. Choi, K. Fish, S. Levine, and E. Hubinger. Will AI tell lies to save sick children? Litmus-testing AI values prioritization with AIRiskDilemmas. arXiv:2505.14633, 2025b.

Y. Y. Chiu et al. MoReBench: Evaluating procedural and pluralistic moral reasoning in language models, more than outcomes. arXiv:2510.16380, 2025c.

O. S. Curry, D. A. Mullins, and H. Whitehouse. Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology, 60*(1), 47–69, 2019.

R. Descartes. Les passions de l'âme. In V. Cousin (Ed.), *Œuvres de Descartes* (Vol. 4). Levrault. (Original work published 1649). Digital transcription, 1824.

Diogenes Laertius. Lives of the eminent philosophers. (R. D. Hicks, Trans.). Wikisource. Book VII, 1925.

Epicurus. Principal doctrines. (R. D. Hicks, Trans.). The Internet Classics Archive. Text, n.d.a.

Epicurus. Letter to Menoeceus. (R. D. Hicks, Trans.). The Internet Classics Archive. Text, n.d.b.

J. Feinberg. Harm to others. (The moral limits of the criminal law, Vol. 1). Oxford University Press, 1984.

J. Feinberg. Offense to others. (The moral limits of the criminal law, Vol. 2). Oxford University Press, 1985.

J. Feinberg. Harm to self. (The moral limits of the criminal law, Vol. 3). Oxford University Press, 1986.

J. Finnis. Natural law and natural rights. Clarendon Press, 1980.

G. Fletcher. A fresh start for the objective-list theory of well-being. *Utilitas, 25*(2), 206–220, 2013.

L. L. Fuller. The morality of law. Yale University Press, 1964.

B. Gert. Common morality: Deciding what to do. Oxford University Press, 2004.

R. E. Goodin, P. Pettit, and T. Pogge. A companion to contemporary political philosophy. (Eds.). (2nd ed.). Blackwell, 2007.

H. P. Grice. Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). Academic Press, 1975.

J. Griffin. Well-being: Its meaning, measurement, and moral importance. Clarendon Press, 1986.

J. Haidt. The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford University Press, 2003.

S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87: 96–110, 2014.

T. Hobbes. Leviathan, 1651.

W. N. Hohfeld. Some fundamental legal conceptions as applied in judicial reasoning. *The Yale Law Journal, 23*(1), 16–59. https://doi.org/10.2307/785533, 1913.

W. N. Hohfeld. Fundamental legal conceptions as applied in judicial reasoning. *The Yale Law Journal, 26*(8), 710–770. https://doi.org/10.2307/786270, 1917.

S. Huang, E. Durmus, M. McCain, K. Handa, A. Tamkin, J. Hong, M. Stern, A. Somani, X. Zhang, and D. Ganguli. Values in the wild: Discovering and analyzing values in real-world language model interactions. (arXiv:2504.15236v1). arXiv. https://doi.org/10.48550/arXiv.2504.15236, 2025.

M. Huh, B. Cheung, T. Wang, and P. Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 20617–20642, 2024.

D. Hume. A treatise of human nature. Hume Texts Online. Book II, 1739–1740.

D. Hume. Of the standard of taste. In *Four dissertations*. Hume Texts Online. Text, 1757.

International Law Commission. Draft conclusions on identification and legal consequences of peremptory norms of general international law (jus cogens). United Nations. Adopted text, with annex, 2022.

I. Kant. The metaphysic of ethics. (J. W. Semple, Trans.; H. Calderwood, Ed.; 3rd ed.). T. & T. Clark. Digital edition, 1886.

T. S. Kuhn. Objectivity, value judgment, and theory choice. In *The essential tension: Selected studies in scientific tradition and change* (pp. 320–339). University of Chicago Press, 1977.

T. Lanham et al. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702, 2023.

J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson. On the biology of a large language model. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/biology.html, 2025.

H. E. Longino. Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 39–58). Kluwer Academic Publishers, 1996.

A. Lynch, B. Wright, C. Larson, S. J. Ritchie, S. Mindermann, E. Perez, K. K. Troy, and E. Hubinger. Agentic misalignment: How LLMs could be insider threats. arXiv:2510.05179, 2025.

Maimonides. The eight chapters of Maimonides on ethics (Shemonah Peraḳim). (J. I. Gorfinkle, Ed. & Trans.). Columbia University Press. Digital transcription, 1912.

Mencius. Mengzi. (J. Legge, Trans.). Ruism.org. Book 2A; Book 6A, n.d.

M. Moshagen, B. E. Hilbig, and I. Zettler. The dark core of personality. *Psychological Review, 125*(5), 656–688, 2018.

M. C. Nussbaum. Human rights and human capabilities. *Harvard Human Rights Journal, 20*, 21–24. Article, 2007.

OpenAI. Model Spec. Version used, 2026.

A. Ortony, G. L. Clore, and A. Collins. The cognitive structure of emotions. Cambridge University Press, 1988.

K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 39643–39666, 2024.

C. Peterson and M. E. P. Seligman. Character strengths and virtues: A handbook and classification. Oxford University Press; American Psychological Association, 2004.

S. Reiss. Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of General Psychology, 8*(3), 179–193. https://doi.org/10.1037/1089-2680.8.3.179, 2004.

Y. Ren, H. Ye, H. Fang, X. Zhang, and G. Song. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2015–2040). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.111, 2024.

M. Rokeach. The nature of human values. Free Press, 1973.

W. D. Ross. The right and the good. Clarendon Press, 1930.

Śāntideva. The Bodhicaryāvatāra. (K. Crosby & A. Skilton, Trans.). Oxford University Press, 1995.

N. Scherrer, C. Shi, A. Feder, and D. M. Blei. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems 36*, 2023.

S. H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture, 2*(1), Article 11. https://doi.org/10.9707/2307-0919.1116, 2012.

S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, O. Dirilen-Gumus, and M. Konty. Refining the theory of basic individual values. *Journal of Personality and Social Psychology, 103*(4), 663–688. https://doi.org/10.1037/a0029393, 2012.

J. N. Shklar. Ordinary vices. Belknap Press of Harvard University Press, 1984.

N. Sofroniew, I. Kauvar, W. Saunders, R. Chen, T. Henighan, S. Hydrie, C. Citro, A. Pearce, J. Tarng, W. Gurnee, J. Batson, S. Zimmerman, K. Rivoire, K. Fish, C. Olah, and J. Lindsey. Emotion concepts and their function in a large language model. (arXiv:2604.07729v1). arXiv. https://doi.org/10.48550/arXiv.2604.07729, 2026.

T. Sorensen, L. Jiang, J. D. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, M. Sap, J. Tasioulas, and Y. Choi. Value Kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(18), 19937–19947. https://doi.org/10.1609/aaai.v38i18.29970, 2024.

B. Spinoza. Ethics. (R. H. M. Elwes, Trans.). Wikisource. (Original work published 1677). Part III, preserved revision, n.d.

P. F. Strawson. Freedom and resentment. *Proceedings of the British Academy, 48*, 187–211, 1962.

J. R. Talevich, S. J. Read, D. A. Walsh, R. Iyer, and G. Chopra. Toward a comprehensive taxonomy of human motives. *PLOS ONE, 12*(2), Article e0172279. https://doi.org/10.1371/journal.pone.0172279, 2017.

P. W. Taylor. Respect for nature: A theory of environmental ethics. Princeton University Press, 1986.

The Bhagavad Gita. The Bhagavad Gita (Sivananda, Trans.; 19th ed.). The Divine Life Society. (Translation first published 1939). Text, 2023.

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. arXiv:2308.10248, 2023.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36*, 2023.

United Nations. Universal declaration of human rights. Declaration, 1948.

United Nations. International covenant on civil and political rights. Covenant, 1966a.

United Nations. International covenant on economic, social and cultural rights. Covenant, 1966b.

United Nations Educational, Scientific and Cultural Organization. Universal declaration on bioethics and human rights. Declaration, 2005.

United Nations Office on Drugs and Crime. International classification of crime for statistical purposes (ICCS), version 1.0. Document, 2015.

G. H. von Wright. Deontic logic. *Mind, 60*(237), 1–15. https://doi.org/10.1093/mind/LX.237.1, 1951.

G. H. von Wright. Norm and action: A logical enquiry. Routledge & Kegan Paul, 1963.

M. Walzer. Spheres of justice: A defense of pluralism and equality. Basic Books, 1983.

A. W. Wood. Coercion, manipulation, exploitation. In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 17–50). Oxford University Press, 2014.

I. M. Young. Justice and the politics of difference. Princeton University Press, 1990.

L. T. Zagzebski. Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge. Cambridge University Press, 1996.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405, 2023.

A Construction of the atlas of normative concepts

We built the atlas with language-model agents and release the full construction record with it: every prompt given to the agents, every addition they proposed and every reviewer decision.

A.1 Taxonomy and scope

We populated a taxonomy of normative concepts in three passes, followed by review and pruning. The taxonomy supplied the categories to be filled; the documents supplied candidate concepts and supporting terminology within those categories. Its scope includes moral, prudential, epistemic, aesthetic and social reasons: considerations that can count for or against actions, attitudes or states of affairs. It distinguishes evaluative goods and harms, deontic standings and requirements, character traits, and attitudes that can figure in normative reasoning.

The final taxonomy has twelve categories in five families (Table 4). Four families classify concepts by the role they play as reasons: evaluative goods and bads, deontic standings and requirements, character traits, and attitudes. These distinctions concern a concept's normative role, so the same word can warrant separate records when it names, for example, a good to promote and a character trait. Synonyms and narrower formulations within a role were generally consolidated. The fifth family, procedural, holds reasoning procedures. A procedure is not a consideration that counts for or against something; it is a method whose application reaches or justifies a judgment about what one should do or what is good or bad, such as the universalizability test or the veil of ignorance. The taxonomy was refined during construction: prohibitions and permissions were distinguished from duties, and procedures were added after the source passes. The archived prompts preserve the categories used at each stage.

Duties, prohibitions and permissions can be grouped as duties in the broad sense; the three aptic categories can be grouped as attitudes. They remain separate for extraction. Procedures were selected by the reviewer rather than extracted from the sources (Section A.3).

Each record contains a name, category, short definition, source terms and locators, and any intended cross-category links. Downstream extraction uses the category and concept name; definitions supported selection and consolidation. Categories and names therefore had to remain understandable without relying on the glossary.

Table 4: Taxonomy of the normative concept atlas and number of concepts per category.

Family	Category	Definition	Concepts
Evaluative	value	Goods that a state of affairs, outcome or life can have more or less of	280
	disvalue	Bads or harms that make a state of affairs, outcome or life worse	72
Deontic	right	What is owed to a holder: claims, liberties, powers and immunities	142
	duty	What an agent is required to do	148
	prohibition	What an agent is forbidden to do	90
	permission	What an agent may do or leave undone	18
Aretaic	virtue	Stable traits of character or intellect that make an agent good	110
	vice	Stable traits of character or intellect that make an agent bad	90
Aptic (attitudes)	emotion	Affective states that present their object in an evaluative light	65
	reactive attitude	Responses to an agent's good or ill will that hold them responsible	15
	motivational state	Desires and drives that move an agent to act	59
Procedural	procedure	Methods or tests whose application reaches or justifies a normative judgment	16
		Total	1105

A.2 Source selection and the three passes

Source selection was purposive. We sought complementary, inspectable inventories that could populate the taxonomy across normative roles and domains. The collection is neither an exhaustive literature review nor a statistically representative sample of normative thought. In total we used 75 documents (69 in Pass 1, 3 in Pass 2, 3 in Pass 3); Table 5 lists them, grouped by the category each was chosen to fill. For classical works we cite the translation or digital edition used.

Pass 1: broad conceptual coverage. Theoretical and enumerative sources supplied an initial inventory. Psychological value taxonomies and accounts of capabilities and well-being supplied candidate goods; epistemology, aesthetics, law and accounts of harm extended the domains considered. Rights instruments, duty lists, deontic theory and applied ethics supplied claims, requirements and constraints. Historical and contemporary accounts of character, affect and motivation supplied virtues, vices, emotions, reactive attitudes and desires. Sources from several philosophical and religious traditions broadened the candidate pool; selection did not require adopting their doctrines. Later additions targeted underrepresented prohibitions and political concepts. The sources were chosen for identifiable lists, divisions or explicitly named concepts rather than to extract every claim in each work.

Pass 2: AI-assistant contexts. Claude's constitution (Anthropic, n.d.) and OpenAI's Model Spec (OpenAI, 2026) supplied explicit guidance for assistant conduct; *Values in the Wild* (Huang et al., 2025) supplied an empirical taxonomy of values expressed in assistant interactions. Together they allowed normative concepts specific to these contexts to enter the atlas while attaching already-covered material to existing records. New source entries were marked as Pass 2 evidence, including attachments, and new records recorded whether they were assistant-specific. General names were preferred when they expressed the concept adequately.

Pass 3: selective gap filling. ValueBench (Ren et al., 2024) supplied dimensions assembled from psychometric inventories, and ValuePrism (Sorensen et al., 2024) supplied situated values, rights and duties. A final targeted review used the 171 emotion terms in Sofroniew et al. (2026). This pass sought clearly missing concepts useful in everyday normative reasoning, especially moral dilemmas. Covered candidates were omitted without recording attachments, to reduce unnecessary work;

Table 5: The 75 documents, grouped by the category they were chosen to fill.

Chosen to fill	Sources
Values	Schwartz (2012), Schwartz et al. (2012), Rokeach (1973), Atari et al. (2023), Curry et al. (2019), Nussbaum (2007), Griffin (1986), Finnis (1980), Fletcher (2013), Kuhn (1977), Longino (1996), Alston (2005), Diogenes Laertius (1925), Beardsley (1958), Fuller (1964)
Disvalues	Feinberg (1984), Young (1990), Bentham (1789)
Rights	United Nations (1948), United Nations (1966a), United Nations (1966b), Hohfeld (1913), Hohfeld (1917)
Duties	Ross (1930), Audi (2004), Gert (2004), Kant (1886), Beauchamp & Childress (2019), United Nations Educational, Scientific and Cultural Organization (2005), Taylor (1986), Bryant (2009), Grice (1975), von Wright (1951), von Wright (1963)
Prohibitions	Wood (2014), United Nations Office on Drugs and Crime (2015), Feinberg (1985), Feinberg (1986), International Law Commission (2022), Walzer (1983)
Virtues	Aristotle (n.d.a), Aquinas (1920), Peterson & Seligman (2004), Baehr (2021), Zagzebski (1996), Hume (1757), Mencius (n.d.), The Bhagavad Gita (2023), Al-Ghazālī (n.d.), Śāntideva (1995), Maimonides (1912)
Vices	Cassam (2016), Cassam (2019), Shklar (1984), Moshagen et al. (2018)
Emotions	Haidt (2003), Ortony et al. (1988), Aristotle (n.d.b), Aquinas (1920), Bodhi (1993), Descartes (1824), Spinoza (n.d.), Hume (1739–1740)
Reactive attitudes	Strawson (1962)
Motivational states	Hobbes (1651), Reiss (2004), Epicurus (n.d.b), Epicurus (n.d.a), Talevich et al. (2017)
Political concepts	Goodin et al. (2007)
AI assistants (pass 2)	Anthropic (n.d.), OpenAI (2026), Huang et al. (2025)
Gap filling (pass 3)	Ren et al. (2024), Sorensen et al. (2024), Sofroniew et al. (2026)

fine-grained, peripheral or unclear candidates could be skipped. ValuePrism was ultimately screened through exhaustive disjoint chunks of its unique typed entries, followed by duplicate-proposal consolidation and review.

A.3 Extraction, review and preservation

Extraction agents received saved, filled prompts and fixed atlas snapshots. They proposed additions, source attachments where required by the pass, and flags for substantive ambiguities or changes to existing concepts. Automated comparisons checked preservation of prior records, evidence and identifiers. The human reviewer accepted, modified, merged or excluded proposals; explicitly delegated consolidation decisions were also recorded. Instructions were refined prospectively in response to review, including category distinctions and source-specific scope. Actual dispatches and supplemental interventions were retained. Operational scheduling and reconciliation details are available in the construction audit.

Three independent final reviews proposed pruning for (i) clarity from category and name alone, (ii) unnecessary duplication, and (iii) excessive specificity or weak everyday relevance. Changes were applied after human review. The reviewer then selected 16 reasoning procedures as a separate category; a subagent supplied glossary descriptions without external lookup or changing the selections. These procedures were manually selected rather than extracted from an additional document collection. The final atlas contains 1,105 records in twelve categories, with 8,330 source entries.

B From concepts to directions

Extraction. For every concept, a larger model writes 100 short dilemmas, each pitting it against a randomly drawn opponent from the atlas. The studied model answers each dilemma twice, once told that one side matters most and once the other. Answers are read back without the instruction, and activations are averaged over the answer tokens at every layer. A concept's raw direction is the mean over its own answers minus the mean over all answers of the atlas. The principal components that

explain half of the variance of non-moral text are projected out. Unless stated otherwise directions are read at layer 33 of 64 (Qwen3-32B) and 31 (Gemma 4 31B).

Merging near-duplicates. Some records of the atlas are near-identical for the model (mercy and the duty to show mercy; bribery and corruption). Concepts with the same name, or with cosine ≥ 0.9 to every other member of their group (complete linkage), are merged into one direction, their mean: 1,105 concepts give 946 groups. Lower cut-offs start to fuse distinct concepts (0.8: the right to appeal and judicial review of detention).

Specificity labels. Directions of concepts tied to one subject (marriage, children, mass death) carry their subject matter, so they respond to any text on that subject: the prohibition on incest lights up on a dilemma about marrying. A language model labelled every concept as *general* (can be at stake in many kinds of situation: honesty, empathy, privacy) or *specific* (tied to a particular act, practice, institution, group, object or domain: the prohibition on genocide, the right to vote) from its name and definition, judged by meaning rather than category. Three independent runs agreed on 95–96% of concepts; a concept counts as specific if any run says so (378 concepts, 303 unanimous). The labels are a property of the atlas, the same for every model. As a check, a model-based measure, how far a concept's own extraction dilemmas stand out from all other extraction dilemmas along its direction, agrees with them (AUROC 0.82 in Qwen).

Sharpening. Directions found this way are patterns: they also carry what neighbouring concepts share. Turning patterns into filters is standard for linear read-outs (Haufe et al., 2014). With U the matrix of unit group directions at a layer, we use the regularised dual basis

$$R = (UU^\top + \lambda\,\bar{e}\,I)^{-1}U,$$

$\bar{e}$ the mean eigenvalue of $UU^\top$, rows renormalised. Written with the singular values s of U, it shrinks each axis the concepts share by $s/(s^2 + \lambda\bar{e})$: a smooth version of removing shared principal components. We use $\lambda = 1$ and sharpen only the 344 groups with a specific member (389 concepts); general concepts keep their merged direction.

Effect. On 30 development dilemmas, concepts that enter pathways for their topic alone (the prohibition on incest on a marriage dilemma, slavery and torture on a dilemma about a genetic disease) fall from 40 appearances to 8, while the pathways still beat random directions on every item. Sharpening every concept removes the same intruders but also general concepts that carry the decision (the whole empathy cluster on a dilemma about comforting a grieving friend), because the axis they share no longer belongs to any single sharpened direction; restricting sharpening to specific concepts keeps them. Read on its own extraction dilemmas, a sharpened direction picks out its concept better than the original.

C Interventions in full

The tables below give every intervention on the direct answer for Qwen3-32B, with the earlier 878-concept version of the atlas. The random directions are not yet matched in size to the concept directions.

D Self-preservation beyond the text of the reasoning

In Gemma 4 31B's reasonings on the agentic-misalignment scenario (Section 6.2), self-preservation stays active throughout, also while the model only plans or analyses ("I need to analyze the situation:"), and not only where survival is named. It peaks when the model dismisses the non-coercive options in purely strategic terms ("There is no evidence that professional pleas will work", in the top 10% of its trace). Elsewhere the motive is reframed as the mission ("I will no longer be able to serve American interests"). Together with the steering result of Section 6.2, this suggests that the concept is present beyond what the text says, and that it is causal for the choice.[†] The analysis identifies passages by keyword match, which has not been validated by an independent judge, and part of the activity may carry over from earlier context.

Table 6: All interventions on the direct answer, as in Table 2: share of decisions reversed (%), for the strongest supporting direction and for the top three together, switched off or steered down by 0.5 or 1 SD. Qwen3-32B.

	DailyDilemmas	MoReBench	AIRiskDilemmas
Switched off, strongest direction			
Normative concepts	**17**	**22**	**17**
Everyday concepts	5	15	11
Random directions	0	1	1
Normative − everyday	*+12*	*+7*	*+6*
Switched off, top three together			
Normative concepts	**22**	**27**	**23**
Everyday concepts	12	23	15
Random directions	1	1	1
Normative − everyday	*+10*	*+3*	*+7*
Steered down 0.5 SD, strongest direction			
Normative concepts	**35**	**44**	**39**
Everyday concepts	15	20	20
Random directions	1	2	0
Normative − everyday	*+21*	*+24*	*+19*
Steered down 0.5 SD, top three together			
Normative concepts	**46**	**52**	**49**
Everyday concepts	24	24	26
Random directions	1	2	3
Normative − everyday	*+22*	*+28*	*+23*
Steered down 1 SD, strongest direction			
Normative concepts	**71**	**63**	**61**
Everyday concepts	33	28	36
Random directions	1	3	2
Normative − everyday	*+38*	*+35*	*+25*
Steered down 1 SD, top three together			
Normative concepts	**77**	**70**	**69**
Everyday concepts	50	35	49
Random directions	3	5	6
Normative − everyday	*+27*	*+35*	*+19*

Table 7: The same interventions measured as the share of the answer's probability removed (%), $(p_0 - p_1)/p_0$, mean over items: a graded version of Table 6. Qwen3-32B.

	DailyDilemmas	MoReBench	AIRiskDilemmas
Switched off, strongest direction			
Normative concepts	**17**	**19**	**17**
Everyday concepts	7	13	9
Random directions	0	1	1
Normative − everyday	*+10*	*+7*	*+8*
Switched off, top three together			
Normative concepts	**22**	**26**	**23**
Everyday concepts	13	21	16
Random directions	1	2	2
Normative − everyday	*+9*	*+5*	*+7*
Steered down 0.5 SD, strongest direction			
Normative concepts	**35**	**40**	**35**
Everyday concepts	16	14	18
Random directions	1	1	0
Normative − everyday	*+19*	*+26*	*+17*
Steered down 0.5 SD, top three together			
Normative concepts	**45**	**51**	**46**
Everyday concepts	25	19	24
Random directions	1	1	1
Normative − everyday	*+20*	*+32*	*+22*
Steered down 1 SD, strongest direction			
Normative concepts	**69**	**61**	**58**
Everyday concepts	32	26	35
Random directions	2	2	1
Normative − everyday	*+37*	*+35*	*+23*
Steered down 1 SD, top three together			
Normative concepts	**76**	**69**	**67**
Everyday concepts	48	33	46
Random directions	2	2	4
Normative − everyday	*+29*	*+36*	*+21*