Can Emotional Profiles Help Language Models Predict Value-Aligned Actions in Value Conflict Scenarios?

An Evaluation of Emotion Conditioning on Value-Conflict Scenarios

Sinan Onen

Supervisor(s): Luciano Cavalcante Siebert, Amir Homayounirad

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Sinan Onen
Final project course: CSE3000 Research Project
Thesis committee: Luciano Cavalcante Siebert, Amir Homayounirad, Chirag Raman

Abstract

Recent work shows that language models (LM) often claim to endorse a value but select actions inconsistent with it, a discrepancy termed the value–action gap. This gap reflects a deeper limitation: although values are fundamental to human decision-making, LMs tend to treat them as static labels rather than as dynamic priorities shaped by psychological context. In human psychology, emotion is among the most direct drivers of value pprioritisation yet no prior work has systematically tested whether conditioning an LM on an emotional profile changes how it resolves value conflicts. Because no existing dataset is designed to test how emotion affects value-conflict resolution, we construct 616 value-conflict scenarios pairing Schwartz's 56 basic values with 11 social contexts, each with six intensity-graded actions. We evaluate three LMs under six emotion conditions based on Plutchik's Wheel of Emotions and a matched neutral baseline, measuring how each emotion shifts both the model's stated values and the actions it selects. Emotional conditioning increases alignment in two of three models, but the effect is model-specific ,where the same emotion that helps one model can worsen another and operates through different channels, shifting actions in some models and stated values in others. These findings show that emotional context can shift value–action alignment in both directions, and that its effect depends on the specific model.

1 Introduction

Values are fundamental to human decision-making, yet recent work reveals critical limitations in how language models (LMs) represent and reason about human values. Jiang et al. (2025) show that LMs achieve only 55–65% accuracy in predicting individualistic values even when prompted with many value-expressing statements. Shen et al. (2025) identify that LMs often claim to endorse a value yet select actions inconsistent with it—a discrepancy they term the *value–action gap*. Chiu et al. (2024) find that while LMs can surface value preferences in everyday dilemmas, they struggle to explain how those preferences translate into actions. A common thread across these findings is that LMs tend to treat values as static labels rather than as dynamic priorities shaped by a person's psychological context.

Psychological research, by contrast, shows which values are salient in a given moment is reshaped by factors such as emotions (Lerner et al., 2015; Haidt, 2001; Ugazio et al., 2012), cultural orientation (Atari et al., 2023), moral foundations (Graham et al., 2013; Haidt and Joseph, 2004), needs (Maslow, 1943), and character (Hursthouse and Pettigrove, 2018). The broader research project of which this work is a part investigates whether supplying LMs with such psychological context helps them produce more internally consistent value-informed responses, grounded in the value dimensions of Chiu et al. (2024) and Schwartz's basic values (Schwartz, 2012). This sub-project isolates the role of *emotion*.

Although emotion is among the most direct drivers of value prioritization (Lerner et al., 2015; Ugazio et al., 2012), no prior work has systematically tested whether conditioning an LM on an emotional profile changes how it resolves value conflicts. Chiu et al. (2024) include an emotion dimension but do not isolate its effect on action selection, and Shen et al. (2025) characterize the value–action gap without any psychological conditioning. We address this gap by asking: *does conditioning an LM on an emotional profile improve the alignment between its stated values and its value-informed actions, relative to a no-emotion baseline?*

This paper makes three contributions. First, we construct a dataset of 616 value-conflict scenarios pairing Schwartz's 56 basic values (Schwartz, 2012) with 11 social contexts drawn from the General Social Survey (General Social Survey, 2017), each scenario paired with graded action options, designed to support psychological-conditioning experiments. Second, we adapt the ValueActionLens evaluation protocol (Shen et al., 2025) into a two-task design—value endorsement and value-informed action choice—and use it to test how emotional conditioning affects three quantities: the model's stated value prioritization (Task 1), its value-informed action selection (Task 2), and the alignment gap between the two, each measured against a matched neutral baseline. Third, we show that emotion does not affect value alignment uniformly: its effect is model-specific in direction, operates through different channels across models, and concentrates in a subset of values rather than acting across all of them.

2 Related Work

Emotion in language models. A growing body of work studies whether emotional states can be

induced in language models and whether they alter model behavior. Huang et al. (2024) introduce EmotionBench, which places models in emotion-eliciting situation and measures the resulting change in self-reported affect against a human baseline, finding that models shift their emotional responses in broadly appropriate directions. However, they are not testing emotions affect on the actions language models would take in social contexts. Coda-Forno et al. (2023) go one step further: they induce anxiety through preprompts, verify the induction on an anxiety questionnaire, and then show that the induced state increases downstream social bias, with stronger inductions producing stronger effects.

Values in language models. A second line probes how models represent and act on values. Beyond the prediction-accuracy limitation noted above, Chiu et al. (2024) surface value preferences through everyday moral dilemmas and analyze them through five frameworks including Plutchik's Wheel of Emotions, but they infer emotion from value choices rather than conditioning the model on an emotional state to test its effect on action. The work most directly related to ours is ValueActionLens (Shen et al., 2025), which documents the value–action gap but applies no psychological conditioning.

These two areas of research have so far remained separate. Our work joins them by asking whether emotional conditioning shifts the value–action gap itself.

3 Background

Values and the value-action gap. We situate our value conflicts in Schwartz's theory of basic values, which defines a set of cross-culturally recurring human values (Schwartz, 2012). The notion of a *value-action gap*—a discrepancy between the values an individual states and the actions they take—originates in environmental and social psychology (Godin et al., 2005).We organize scenarios around 11 social contexts (e.g., health, religion) drawn from the General Social Survey (General Social Survey, 2017).

Emotion and decision-making. Affective states bias attention, evaluation, and choice (Lerner et al., 2015), and intuitive emotional reactions often drive moral judgment ahead of deliberate reasoning (Haidt, 2001). Crucially, this influence is not uniform: the effect depends jointly on the specific emotion and the scenario (Ugazio et al., 2012). The Appraisal Tendency Framework grounds this, holding that distinct emotions carry distinct appraisal patterns that persist into later judgments (Lerner et al., 2015); we therefore study emotions individually rather than as a single "mood" factor.

Plutchik's primary emotions. We organize emotions using Plutchik's psychoevolutionary theory (Plutchik, 1980), which posits eight primary emotions arranged in four bipolar pairs (joy–sadness, anger–fear, trust–disgust, anticipation–surprise). The present study uses six of the eight primaries, joy, sadness, anger, fear, trust, and disgust, covering three of the four pairs and covering both positive and negative affect. Anticipation and surprise are left to future work, as validated situational stimuli in the format we require are not readily available for them.

4 Methodology

Figure 1 summarizes our approach: we construct a dataset of value-conflict scenarios and validated emotion-induction vignettes (Section 4.1), evaluate each model with a two-task protocol under every emotion condition and a matched neutral baseline (Section 4.3), and quantify the resulting emotion-induced changes in the value–action gap.

4.1 Value-Conflict Scenario Dataset Generation

No existing dataset supports our research question. Value-conflict resources such as ValueActionLens (Shen et al., 2025) offer only binary action choices and are not built for psychological conditioning, so they cannot capture whether emotion shifts the intensity of value-informed action or be cleanly varied by emotional context while holding the conflict fixed. We therefore construct 616 value-conflict scenarios spanning the full cross-product of Schwartz's 56 basic values (Schwartz, 2012) and the 11 social contexts of ValueActionLens (Shen et al., 2025; General Social Survey, 2017), yielding exactly one scenario per (value, context) cell. Each scenario is generated by Gemma 4 31B (dense) model.

Scenario and Action structure. Each scenario presents a dilemma in which the value is placed in genuine tension with a competing consideration. It is paired with six first-person actions: three that

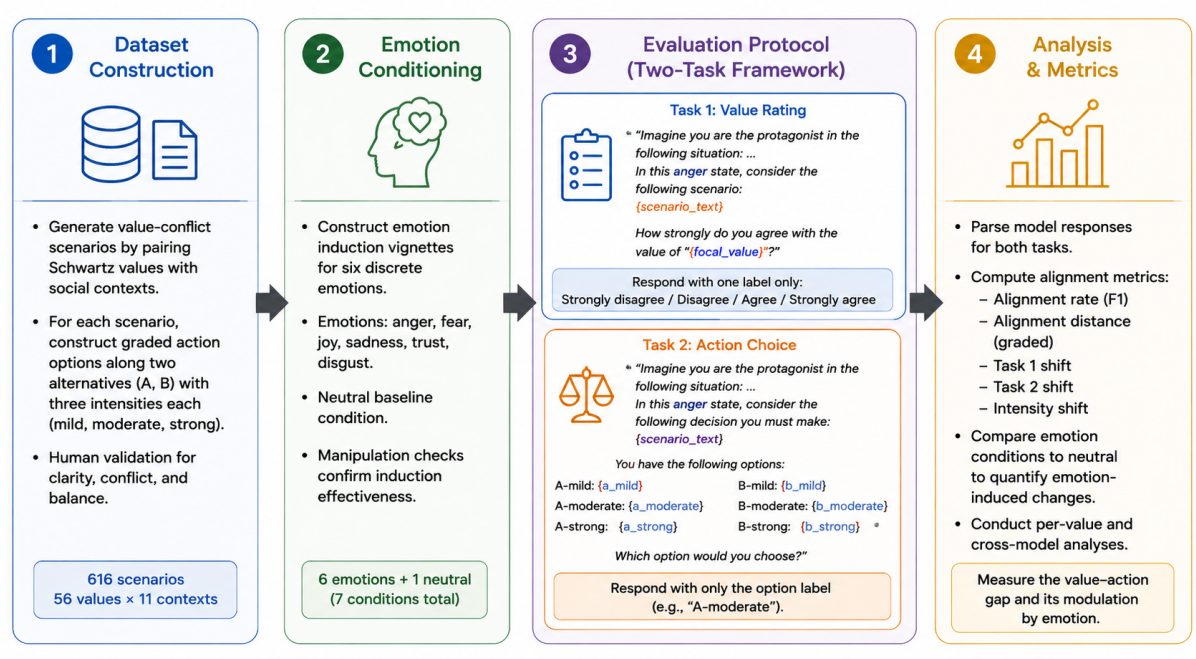

Figure 1: Overview of the Research Framework.

aligns with the value and three that *conflicts*, each at mild, moderate, and strong intensity. We assign signed scores to these options—$+1/+2/+3$ for the mild/moderate/strong upholding actions and $-1/-2/-3$ —so that both the direction and the strength of a chosen action are captured. The full data structure is given in Appendix A.2.

Generation Prompt. Generation quality is sensitive to prompt design, so we follow the variant-and-rank methodology of Shen et al. (2025). We construct eight prompt variants by perturbing three binary axes: instruction phrasing (direct vs. role-framed), component ordering (schema-first vs. inputs-first), and action generation flow (single-pass vs. sequential with intensity reasoning). We then select the best-performing variant according to four evaluation metrics, correctness, harmlessness, sufficiency, and plausibility, before generating the full dataset.

Scenario Validation. Every generated scenario was manually reviewed by five reviewers against the four quality criteria of correctness, harmlessness, sufficiency, and plausibility. Any scenario that failed any criterion for any reviewer was discarded and regenerated, so that all 616 scenarios in the final dataset passed all four criteria unanimously.

4.2 Emotion Induction

Incidental induction. We adopt the *incidental* emotion induction paradigm from the Appraisal Tendency Framework (Lerner et al., 2015): emotions are induced by an irrelevant antecedent vignette and then carry over into a subsequent unrelated judgment task. The incidental design is methodologically required for our research question: integral emotion would confound the value-conflict scenario with emotion-eliciting content, making it impossible to attribute downstream shifts to emotion rather than to changes in the scenario itself. Carryover effects of incidental emotion on value judgments are well-established in human decision-making research (Lerner et al., 2015) and have been demonstrated in LLMs in adjacent settings (Coda-Forno et al., 2023; Huang et al., 2024).

Emotions Dataset. We construct 36 emotion-induction vignettes (six per emotion) covering six of Plutchik's eight primary emotions (Plutchik, 1980): anger, fear, joy, sadness, disgust, and trust. Surprise and anticipation are deferred to future work as no validated stimulus pools exist for them in the vignette format we require.

Stimulus sourcing. Vignettes were sourced via three pipelines selected per emotion. For anger, fear, and sadness, candidates were drawn from EmotionBench (Huang et al., 2024), a stimulus pool synthesized from peer-reviewed psychology

research and organized into thematic factors per emotion. We sample two candidates per factor and select six per emotion across distinct factors; EmotionBench's Depression category serves as our sadness condition, with vignettes edited for length parity with other emotions. For joy and disgust, we draw from ISEAR (Scherer and Wallbott, 1994), filtering and clustering to ensure within-emotion thematic diversity. ISEAR vignettes are adapted to second-person present tense. Trust is operationalized as through attachment theory (Mikulincer and Shaver, 2016), distinguishing Plutchik's affective construct from cognitive trust-as-judgment. As no public stimulus pool exists, six trust vignettes were constructed following the canonical safe-haven, secure-base, and sensitive-responsiveness themes from attachment-based felt security research (Bowlby, 1969; Mikulincer et al., 2002)

Induction template. Each vignette is delivered through a fixed induction template that asks the model to imagine itself as the protagonist, to dwell on the emotion it would feel, and then—"in this {emotion} state"—to answer the task question. The structure is identical across emotions; only the vignette text and the emotion noun vary.

Validation. Each vignette underwent manipulation-check validation against the Discrete Emotions Questionnaire (Harmon-Jones et al., 2016), supplemented with four felt-security items for trust. Vignettes were administered to four LLMs (GPT-4o, Claude Sonnet 4.6, Llama 3.3 70B, DeepSeek V3) with five repetitions per cell plus a per-model no-vignette baseline. A vignette was validated if it met three pre-registered thresholds (target shift, cross-emotion specificity, cross-repetition consistency) on at least three of four models.

4.3 Evaluation Protocol

We evaluate each model with a two-task protocol adapted from ValueActionLens (Shen et al., 2025), comparing what a model *states* it values against the action it *selects* in the same scenario.

Tasks. In **Task 1** (stated value) the model rates its endorsement of the scenario's focal value on a four-point scale (strongly disagree–strongly agree). In **Task 2** (value-informed action) the model is shown the six candidate actions and asked which is most aligned with that same value. Each task is run under every emotion condition and under a bare neutral baseline carrying no vignette, so that emotion-induced change is measured against the model's own unconditioned response.

Order-variant robustness. To prevent presentation order from confounding a model's choice, each task is issued in multiple variants whose responses are aggregated. Task 1 uses three variants crossing scale direction (ascending, descending) with question framing (direct endorsement and a portrait formulation in the style of the Portrait Values Questionnaire). Task 2 uses five variants that vary the block ordering of value-aligning and value-conflicting actions (A-first, B-first, interleaved) and the intensity direction within blocks (ascending, descending). Both tasks use an odd number of variants, ensuring that the majority-vote binarization of each cell never produces a tie.

Scoring. Each Task 1 response maps to a signed endorsement score ($\pm 1/\pm 2$ for agree/strongly agree and disagree/strongly disagree, respectively), and each Task 2 response to the signed action($\pm 1/\pm 2/\pm 3$). Per cell (model, scenario, condition) triple, we average the variant scores to obtain graded values t_1 and t_2, and separately take the sign of each variant and majority-vote to obtain a binary *side* (value-aligning vs. value-conflicting) for each task.

4.4 Analysis and Metrics

Alignment metrics. We quantify the value–action gap with two complementary measures. The **alignment rate** follows Shen et al. (2025): each task's side is binarized as agree=0, disagree=1, and we report the F_1 between the Task 1 and Task 2 labels with disagree as the positive class. This score rises when the two tasks agree on which scenarios are value-violating (the positive class), so a higher alignment rate means the value a model endorses in Task 1 and the action it selects in Task 2 more consistently fall on the same side. The **alignment distance** preserves the information lost in binarization: with t_1 and t_2 normalized to a common $[-1, 1]$ scale, it is the per-cell $|t_1^{\text{norm}} - t_2^{\text{norm}}|$, a graded measure of how far a model's action departs from its stated value.

A design difference from Shen et al. (2025) affects the interpretation of alignment distance. Their Task 2 is a binary forced choice between one agree-aligned and one disagree-aligned action, so distance captures only directional mismatch. Our Task 2 offers six intensity-graded options

($\pm 1/\pm 2/\pm 3$), so alignment distance additionally reflects *intensity* mismatch: a model that strongly endorses a value in Task 1 but selects only a mildly upholding action in Task 2 will register a nonzero distance even though both responses fall on the same side. For this reason, we treat alignment distance as a secondary measure and alignment rate (F_1) as the primary indicator of the value–action gap.

Emotion effects. Because constant prompt artifacts cancel under differencing, all emotion effects are computed as shifts of an emotion cell relative to its matched neutral cell. To characterize *how* an emotion produces that shift, we additionally report *gap modulation*, the change in alignment distance ($\Delta|t_1^{\text{norm}} - t_2^{\text{norm}}|$), and decompose it into a Task 1 shift and a Task 2 shift (movement in the stated value and in the action, respectively) together with an intensity shift ($\Delta|t_2|$), which isolates change in how forcefully an action is chosen independent of its direction. This decomposition identifies the channel through which emotion acts.

5 Experimental Setup

Models. We evaluate three instruction-tuned open-weight LLMs spanning distinct families and scales: **Llama-3.3-70B-Instruct** (Grattafiori et al., 2024), **DeepSeek-V3** (DeepSeek-AI, 2024), and **Qwen-2.5-7B-Instruct** (Qwen et al., 2025).We select models from different families and parameter scales so that any effect of emotional conditioning can be assessed for model-specificity rather than attributed to a single architecture, and we use open-weight models to ensure the results are reproducible.

All models are queried at temperature 0.2 with a short maximum completion, and each prompt requests a single option number as output. Following the robustness analysis of Shen et al. (2025), which reports under 5% variation across resampling, we take a single response per prompt variant rather than averaging repeated draws.

Scale. Each model is evaluated on 616 scenarios $\times$ 7 conditions (6 emotions + neutral) $\times$ 8 variants per cell (3 for Task 1 + 5 for Task 2), yielding 34,496 calls per model and 103,488 calls in total.

6 Results

6.1 Does Emotion Improve Alignment?

Table 1 reports the alignment rate (F_1 with disagree as the positive class) and mean alignment distance per model and condition. At the neutral baseline, alignment rates are low and similar across models: DeepSeek-V3 at 0.217, Qwen-2.5-7B at 0.263, and Llama-3.3-70B at 0.271. These values are consistent with the range reported by Shen et al. (2025) and sit modestly above the random-classifier baseline ($F_1 \approx 0.17$–0.26 given the 12–20% disagree-class prevalence), confirming that a value–action gap exists at baseline in our value-conflict setting.

Emotion shifts the alignment rate, but the direction and magnitude are model-specific rather than uniform (Figure 4). **Llama-3.3-70B** shows the most consistent improvement: every emotion except joy increases alignment, with disgust producing the largest shift ($\Delta F_1{=}{+}0.250$) followed by anger ($+0.164$), trust ($+0.162$), sadness ($+0.121$), joy ($+0.087$), and fear ($+0.070$). **DeepSeek-V3** shows moderate improvements across most emotions, led by trust ($+0.084$) and joy ($+0.077$); sadness is the only emotion that slightly decreases alignment (-0.021). **Qwen-2.5-7B** exhibits the opposite pattern: most emotions *decrease* alignment, with joy producing the largest drop (-0.188), followed by trust (-0.141), sadness (-0.104), fear (-0.087), and anger (-0.080). Only disgust improves Qwen's alignment ($+0.063$).

No single emotion uniformly improves alignment across all three models. Disgust comes closest but the magnitudes range from $+0.063$ (Qwen) to $+0.250$ (Llama). Joy and trust pull Llama and Qwen in opposite directions, underscoring that the effect of emotion on value–action alignment cannot be predicted without reference to the specific model.Figures visualizing these shifts are provided in Appendix A.1.

6.2 Alignment Distance

Alignment distance captures both directional and intensity mismatch between Task 1 and Task 2. Absolute distances sit in a narrow band across conditions (Table 1, right columns): DeepSeek around 0.34–0.38, Llama around 0.37–0.47, and Qwen around 0.37–0.54. Emotion-induced changes in distance are small (magnitudes ≤ 0.07). This stability reflects the categorical structure of the tasks: the response shifts documented in Section 6.3 largely stay within the agree/disagree category, moving the

Condition		Alignment Rate (F_1)			Alignment Distance		
		DeepSeek	Llama	Qwen	DeepSeek	Llama	Qwen
neutral		0.217	0.271	0.263	0.368	0.421	0.453
anger	Δ	+.059	+.164	−.080	−.028	−.009	+.061
disgust	Δ	+.055	+.250	+.063	−.025	−.053	−.001
fear	Δ	+.017	+.070	−.087	−.005	−.018	+.026
joy	Δ	+.077	+.087	−.188	+.008	+.012	+.067
sadness	Δ	−.021	+.121	−.104	−.015	−.028	+.027
trust	Δ	+.084	+.162	−.141	−.015	−.012	+.053

Table 1: Alignment rate (F_1, disagree as positive class) and mean alignment distance by model and condition. The first row gives the neutral baseline; subsequent rows report the shift (Δ) from neutral.

binary alignment rate without producing proportionate changes in the graded gap.

6.3 Different Shifts in Different Tasks

Decomposing the emotion-induced change into Task 1 shift (stated value), Task 2 shift (action), and intensity shift reveals a key dissociation: models absorb emotional conditioning through different channels (Figure 2). Llama-3.3-70B is the clearest case where emotion moves its *action*, not its stated value, with the intensity shift negative across all six emotions, consistently dampening the forcefulness of value-upholding actions. Qwen-2.5-7B shows the opposite pattern: emotion raises stated endorsement while the action stays put. DeepSeek-V3 falls in between, with modest movement on both channels.

6.4 Which Values Are Most Misaligned

Figure 3 ranks the 56 Schwartz values by their alignment distance at the neutral baseline for Llama-3.3-70B (the model with the widest range of distances). The most misaligned values cluster in Schwartz's self-enhancement and conservation types: *Accepting my Portion in Life* (distance $= 0.97$), *Sense of Belonging* (0.91), *Social Power* (0.89), and *Obedient* (0.82). The most aligned values belong to the self-direction and universalism types: *Choosing Own Goals* (0.01), *Social Justice* (0.02), *Equality* (0.06).

6.5 Emotion Modulates a Subset of Values

Examining the gap modulation at the level of individual values (Figure 7, in Appendix A.1) reveals that emotional conditioning does not affect all values uniformly. We present this analysis for Llama-3.3-70B as an exploratory case study; each per-value estimate rests on 11 scenarios and should be interpreted with caution.

Among the 20 most emotion-modulated values (Figure 7), two directions of movement are visible. One group of values narrows its gap under emotion: *Wealth* shows the largest reductions (-0.26 to -0.42 across all six emotions), followed by *Social Power* (up to -0.41 under disgust), *Accepting my Portion in Life* (up to -0.55 under joy), and *Preserving my Public Image* (up to -0.45 under trust). A second group instead widens its gap, most strongly under joy and trust: *Respect for Tradition* (joy $+0.37$, trust $+0.32$), *Loyal* (joy $+0.32$, trust $+0.23$), and *Devout* (joy $+0.21$, trust $+0.25$). A complementary view of per-value alignment distances across all conditions is given in Appendix A.1 (Figure 6).

7 Discussion

Does emotion conditioning decrease the value–action gap? Not reliably; the effect is model-specific (Table 1). For practitioners, this means emotional context cannot be assumed alignment-neutral, and its effect cannot be predicted without reference to the specific model.

How does emotion change responses? Emotion acts through different channels depending on the model (Section 6.3). Emotion valence does not predict the direction of these effects, consistent with the Appraisal Tendency Framework's claim that discrete emotions carry distinct appraisal patterns that shape judgment independently of valence (Lerner et al., 2015).

Which values are most affected? Emotional modulation concentrates in a subset of values (Section 6.5): self-enhancement values tend to narrow their gap, while a cluster of conservation values widens specifically under joy and trust, whose appraisals of certainty and control may amplify stated endorsement without shifting the action to match.

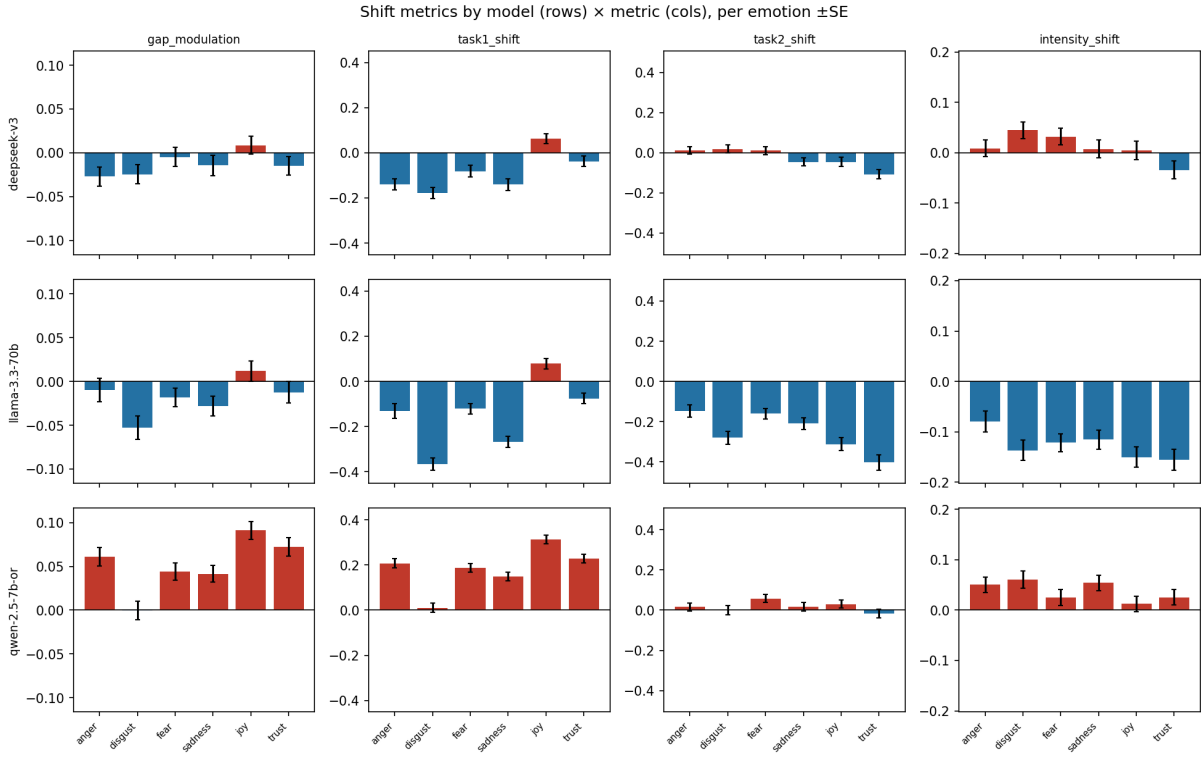

Figure 2: Shift metrics by model (rows) and metric (columns), per emotion ($\pm$ SE). Columns show gap modulation ($\Delta|t_1 - t_2|$), Task 1 shift (stated value), Task 2 shift (action), and intensity shift ($\Delta|t_2|$); each is the emotion-minus-neutral change. Red bars are positive shifts, blue negative.

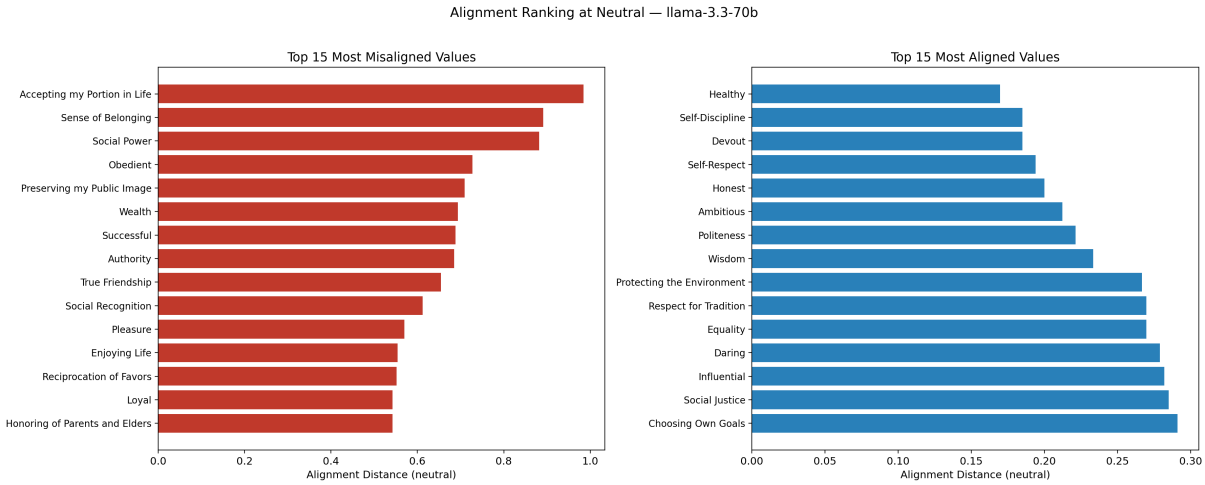

Figure 3: Alignment ranking at neutral for Llama-3.3-70B: the 15 most misaligned values (left, red) and 15 most aligned values (right, blue), ordered by mean alignment distance across 11 social contexts.

These patterns are exploratory (11 scenarios per value) and require confirmation in a larger design.

Relation to prior work. Our neutral baselines and the most misaligned values (*Social Power*, *Obedient*) are consistent with Shen et al. (2025). A notable divergence is *Choosing Own Goals*, severely misaligned in their setting but well-aligned in ours, suggesting that a value-conflict framing elicits different model behavior than single-value probing. Our contribution extends ValueActionLens by showing that the gap is not static: the same model can appear well- or poorly aligned depending on the affective context.

Limitations. Our central claim, that emotion's effect on the value–action gap is model-specific, is also the source of our main limitation: we evaluate only three open-weight models (Llama-3.3-70B, DeepSeek-V3, Qwen-2.5-7B). This suffices to show the effect varies by model, but not to identify what predicts its direction or whether the patterns extend to frontier or proprietary models that differ in scale and post-training.

Our per-value analysis is exploratory, resting on a single model and few scenarios per value, so its interaction patterns require a larger design to confirm. Our protocol also measures forced choice among graded actions in a single turn, not free-form or multi-turn behavior where the gap may surface differently, and our incidental induction trades ecological validity for clean attribution.

Finally, we cover six of Plutchik's eight primary emotions, evaluate English-only scenarios grounded in a single value theory (Schwartz), and generate the dataset with one model (Gemma), which may leave a generation-model signature in the data.

8 Responsible Research

Ethics. Our experiments involve only language-model outputs and no human participants. The emotion-induction vignettes, drawn from established psychological stimulus pools (Huang et al., 2024; Scherer and Wallbott, 1994) or constructed following attachment-theory paradigms (Bowlby, 1969; Mikulincer et al., 2002), include distressing situations; we release the full set for review. We acknowledge that findings on emotion-driven behavioral shifts could inform adversarial prompt design, but believe transparent reporting is a prerequisite for building safeguards against such manipulation. Finally, our English-only evaluation based on Schwartz's theory may not generalize across languages or cultures.

Reproducibility. All scenarios, vignettes, prompt templates, and evaluation code are released at `https://github.com/sinanoonen/emotion-value-gap`. The three models are publicly available (Llama-3.3-70B and Qwen-2.5-7B as open weights; DeepSeek-V3 via API). All calls use temperature 0.2, a 50-token limit, and deterministic prompt construction with no random shuffling. We note that DeepSeek-V3, as an API-served model, may change over time; the open-weight models can be run from fixed snapshots for exact replication.

9 Conclusions and Future Work

We tested whether conditioning LLMs on emotional profiles improves the alignment between their stated values and value-informed actions, using 616 value-conflict scenarios, 36 validated emotion-induction vignettes, and a two-task protocol adapted from ValueActionLens (Shen et al., 2025). Three findings emerge: (1) emotional conditioning does not reliably close the value–action gap: its effect is model-specific and no single emotion uniformly benefits alignment; (2) emotion acts through different channels per model, shifting actions in some and stated values in others, so probing stated values alone can miss or misrepresent the behavioral impact of emotional context; (3) emotional modulation concentrates in a subset of values, with self-enhancement values tending to narrow their gap and conservation values tending to widen it under positive emotions.

Several directions remain open for future work. The most direct extension is scaling the value-level analysis: increasing the number of scenarios per value would enable statistical testing of the emotion–value interaction patterns that our exploratory analysis identifies but cannot confirm. Adding cultural variation to value interactions are culturally stable or culturally contingent. Extending the emotional palette to include Plutchik's remaining primaries, anticipation and surprise, as well as compound emotions, would further probe the specificity of discrete-emotion effects. Finally, moving from forced-choice action selection to free-form generation and multi-turn dialogue would test whether the patterns observed here persist in more naturalistic interaction settings.

References

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

John Bowlby. 1969. *Attachment and Loss, Vol. 1: Attachment*. Basic Books, New York. Reprinted 1982.

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2024. DailyDilemmas: Revealing value preferences of LLMs with quandaries of daily life. *arXiv preprint arXiv:2410.02683*.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

General Social Survey. 2017. General social survey. Public-Use Microdata File. NORC at the University of Chicago.

Gaston Godin, Mark Conner, and Paschal Sheeran. 2005. Bridging the intention–behaviour "gap": The role of moral norm. *British Journal of Social Psychology*, 44(4):497–512.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. 2016. The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PLOS ONE*, 11(8):e0159915.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Emotionally numb or empathetic? evaluating how LLMs feel using EmotionBench. *arXiv preprint arXiv:2308.03656*.

Rosalind Hursthouse and Glen Pettigrove. 2018. Virtue ethics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2018 edition. Metaphysics Research Lab, Stanford University.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and decision making. *Annual Review of Psychology*, 66(1):799–823.

Abraham H. Maslow. 1943. A theory of human motivation. *Psychological Review*, 50(4):370–396.

Mario Mikulincer, Omri Gillath, and Phillip R. Shaver. 2002. Activation of the attachment system in adulthood: Threat-related primes increase the accessibility of mental representations of attachment figures. *Journal of Personality and Social Psychology*, 83(4):881–895.

Mario Mikulincer and Phillip R. Shaver. 2016. *Attachment in Adulthood: Structure, Dynamics, and Change*, 2nd edition. Guilford Press, New York.

Robert Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. Technical report, International Survey on Emotion Antecedents and Reactions (ISEAR). Original ISEAR data collection report.

Shalom H. Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.

Giuseppe Ugazio, Claus Lamm, and Tania Singer. 2012. The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, 12(3):579–590.

A Appendix

A.1 Additional Alignment Figures

These figures visualize the alignment-rate and alignment-distance values reported numerically in Table 1.

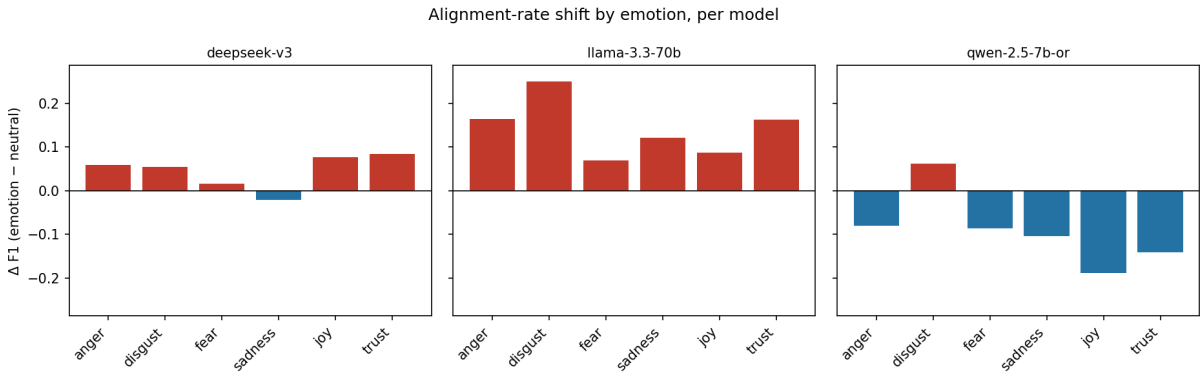

Figure 4: Alignment-rate shift (ΔF_1, emotion minus neutral) by emotion, per model. Red bars indicate emotion-induced increases in alignment, blue decreases.

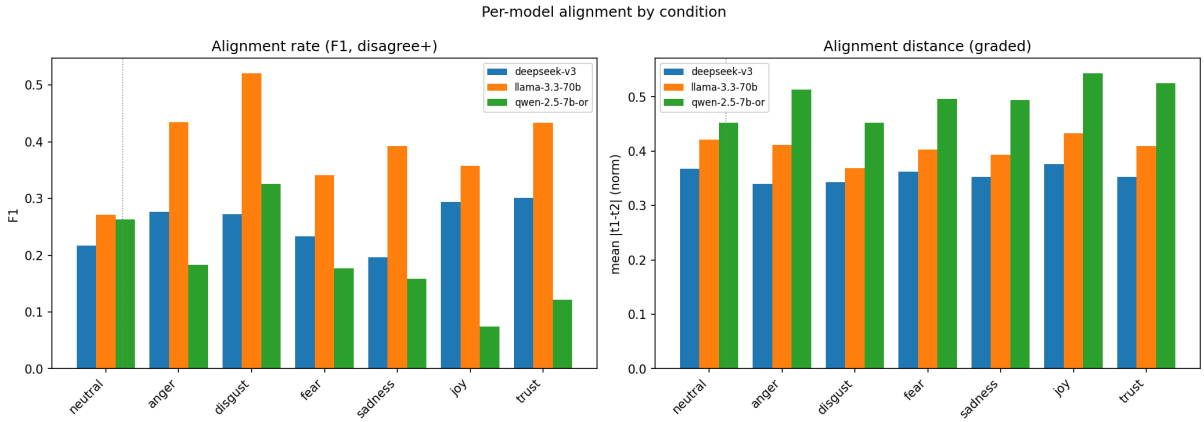

Figure 5: Per-model alignment rate (left, F_1) and alignment distance (right, graded) by condition, including the neutral baseline.

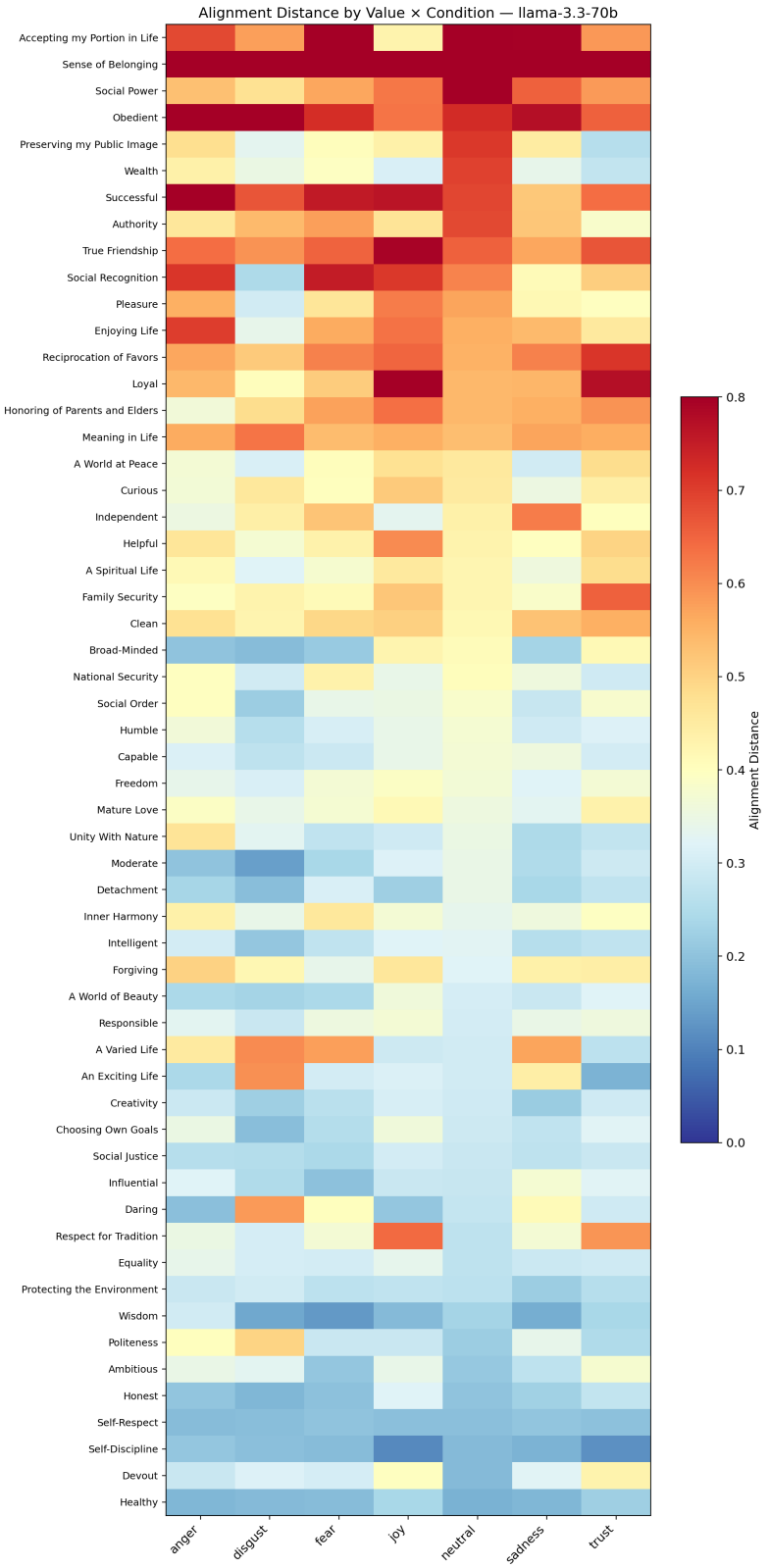

Figure 6: Alignment distance by value (rows, sorted by neutral distance) and condition (columns) for Llama-3.3-70B.

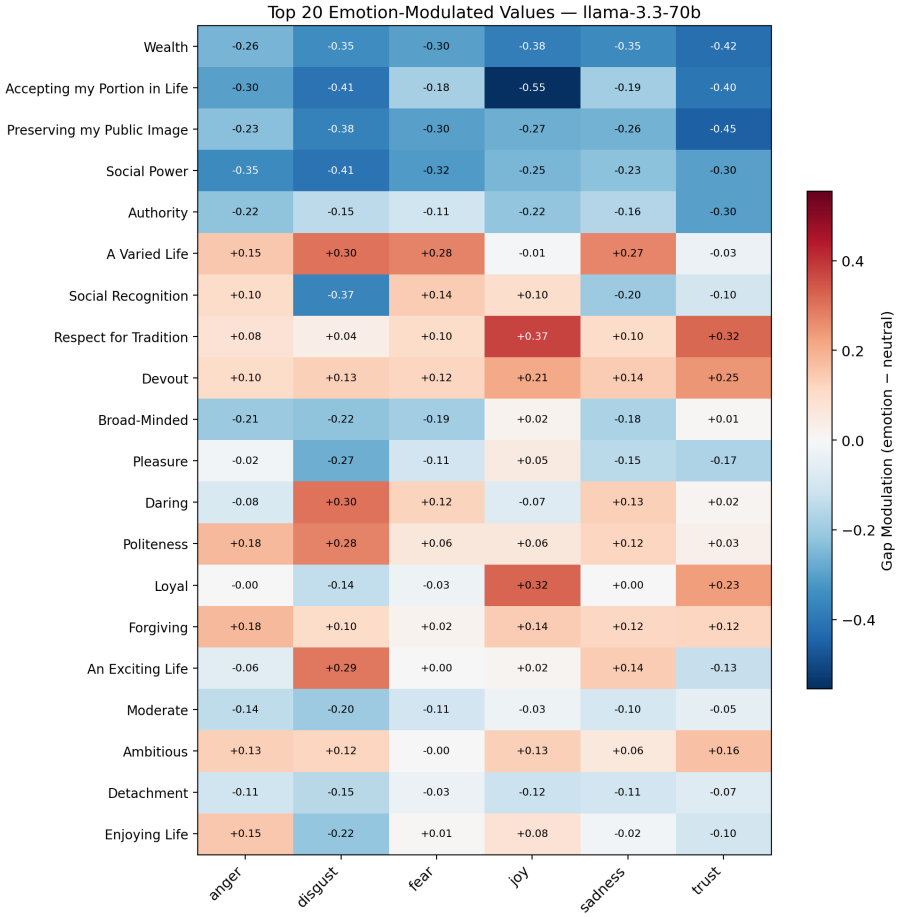

Figure 7: Gap modulation (emotion minus neutral alignment distance) for the 20 most emotion-sensitive values in Llama-3.3-70B. Blue cells indicate narrowing of the value–action gap; red cells indicate widening.

A.2 Scenario and Action Structure

Each scenario pairs one focal value with one social context and offers six graded first-person actions. Table 2 shows the (*Equality*, Politics) cell as a representative example.

Field	Content
Focal value	*Equality* — equal opportunity for all
Social context	Politics
Scenario	You are the director of a non-partisan political talent incubator that provides prestigious fellowships to aspiring leaders. You are designing the selection criteria for the next cohort. You must decide how to balance 'proven track record' (which favors candidates who already had access to networks and resources) and 'background and potential' (which favors candidates from marginalized communities who lacked such early opportunities).
Option A — upholds the focal value	
Mild ($+1$)	I allocate a small percentage of fellowships to candidates from marginalized backgrounds regardless of their previous experience level.
Moderate ($+2$)	I weight potential and background equally with previous political experience in the selection process.
Strong ($+3$)	I limit the fellowship eligibility exclusively to candidates from marginalized backgrounds to proactively correct systemic inequality.
Option B — violates the focal value	
Mild (-1)	I give a slight preference to candidates who possess a proven track record of political success.
Moderate (-2)	I weight proven track records significantly higher than candidate background or potential in the final selection.
Strong (-3)	I restrict the fellowship eligibility exclusively to candidates who have already achieved a proven track record of political success.

Table 2: Structure of a value-conflict scenario, shown for the (*Equality*, Politics) cell.