The Context Sensitivity Paradox: How Stakeholder Framing Shapes Moral Judgment in Humans and AI

Jonathan H. Westover *

Posted Date: 24 February 2026

doi: 10.20944/preprints202602.1434.v1

Keywords: business ethics; moral reasoning; contextual sensitivity; artificial intelligence; organizational decision-making; moral psychology; particularism

 doi:10.20944/preprints202602.1434.v1

Article

The Context Sensitivity Paradox: How Stakeholder Framing Shapes Moral Judgment in Humans and AI

Jonathan H. Westover

Western Governors University, USA; jon.westover@gmail.com

Abstract

Do humans and artificial intelligence systems apply consistent ethical frameworks when making organizational decisions, or do morally contested contextual features systematically influence moral judgments? We conducted a mixed-methods experiment with 300 organizational leaders and three frontier AI models (GPT-4, Claude 3 Opus, Gemini Pro 1.5) responding to 240 systematically varied ethical scenarios. We collected 6,000 human responses (300 participants × 20 scenarios each) and generated 7,200 AI responses at temperature 0.7 (240 scenarios × 3 models × 10 repetitions), with additional sensitivity analyses at temperatures 0.3, 0.5, and 1.0. Analysis revealed substantial systematic variation in moral judgments: both humans and AI systems (at temperature=0.7, selected post-hoc to match human total variation levels) made different recommendations for structurally identical dilemmas based on how stakeholders were described (identifiable individuals vs. statistical aggregates: OR=2.08, p<.001), whether actions were framed as active causation versus passive allowance (d=0.63, p<.001), and temporal proximity of consequences (OR=1.52, p<.001). We decomposed this variation into three components: (1) structural consistency (agreement when only irrelevant features vary: M=0.85), (2) contextual responsiveness (variation attributable to debatable features: 22-24% of total variance), and (3) arbitrary residual variation (32-34% of variance). AI temperature parameter directly controls variation magnitude: at T=0.3, AI showed less variation than humans (0.26 vs. 0.42, p<.001); at T=0.7 (typical deployment setting), AI approximated human levels (0.41 vs. 0.42, p=.56); at T=1.0, AI exceeded human variation (0.49 vs. 0.42, p<.001) but with degraded coherence (91.6% vs. 97.2% at T=0.7). This temperature-dependence means we cannot claim AI inherently exhibits human-like variation; rather, temperature embeds implicit assumptions about desired reasoning patterns. Critically, contextual feature effects (identifiability, relational, temporal, action-omission) remained significant across all temperatures (all $\eta^2p > 0.08$, p<.001), indicating robust patterns independent of overall variation levels. Humans exhibited more relational reasoning than AI (d=0.56, p<.001). Mediation analysis on a coded subsample of 800 responses revealed that relational reasoning partially explained contextual responsiveness differences (indirect effect β=0.043, 95% CI [0.028, 0.061]), accounting for 69% of human-AI differences. Critically, relational reasoning was associated with higher contextual responsiveness but lower arbitrary variation, suggesting systematic sensitivity rather than random inconsistency. Whether contextual responsiveness represents cognitive bias or appropriate moral sensitivity remains philosophically contested. Principlism interprets our findings as evidence of widespread moral reasoning failures; particularism interprets identical patterns as appropriate attention to morally relevant contextual details. Our data provide empirical constraints for this normative debate but cannot adjudicate it.

Keywords: business ethics; moral reasoning; contextual sensitivity; artificial intelligence; organizational decision-making; moral psychology; particularism

 doi:10.20944/preprints202602.1434.v1

1. Introduction

1.1. The Consistency Paradox in Organizational Ethics

When organizational leaders face ethical dilemmas—whether to lay off employees, how to allocate limited resources, or when to prioritize stakeholder groups—should identical decisions follow from structurally identical situations? A principled approach to business ethics suggests "yes": moral reasoning requires identifying universal principles that apply consistently regardless of surface features (Beauchamp & Childress, 2019; Kant, 1785/1993). Yet extensive research in moral psychology demonstrates that superficial variations in how dilemmas are presented—whether victims are identified by name or described statistically, whether harmful outcomes result from action or inaction, whether consequences are immediate or delayed—substantially alter moral judgments (Greene, 2013; Small et al., 2007; Sunstein, 2005).

This creates a paradox: **The same empirical patterns appear as reasoning failures through one philosophical lens and as appropriate moral sensitivity through another.** Consider a corporate restructuring scenario. If an executive recommends layoffs when stakeholders are described as "50 employees in Division A" but resists identical layoffs when those same individuals are named "Maria Rodriguez and 49 colleagues with an average tenure of 12 years," has the executive demonstrated:

(a) **Cognitive bias** (identifiable victim effect undermining impartial analysis), or
(b) **Moral sensitivity** (appropriate recognition of individual dignity versus statistical abstraction)?

Traditional approaches to business ethics assume (a)—that effective ethical decision-making requires transcending contextual particulars to identify universal principles (Freeman, 1984; Rawls, 1971; Singer, 1972). By this standard, systematic variation across presentations of structurally identical scenarios constitutes moral reasoning failure. Yet alternative philosophical traditions argue for (b)—that contextual details including stakeholder identities, relational contexts, and specific circumstances constitute morally relevant information that should shape judgment (Dancy, 2004; Gilligan, 1982; Levinas, 1969).

1.2. The Empirical-Normative Gap

This paradox reflects a deeper philosophical tension between **principlism** and **particularism** in moral philosophy:

Principlism holds that moral reasoning requires identifying universal principles—rules that apply consistently regardless of how situations are described or who is affected. From this view:

- Allowing stakeholder identifiability to alter recommendations violates impartiality (treating like cases alike)
- Action/omission distinctions represent irrational omission bias rather than morally significant differences
- Temporal proximity effects reflect present bias and hyperbolic discounting
- Relationship-based reasoning introduces arbitrary favoritism

Surface feature effects thus represent **cognitive biases** that undermine ethical reasoning and should be minimized through training, institutional safeguards, and deliberative procedures (Bazerman & Tenbrunsel, 2011; Kahneman, 2011).

Particularism counters that morally relevant features cannot be separated from contextual details. From this view:

- Responding differently to named individuals versus statistical abstractions reflects appropriate attention to persons rather than irrational bias
- Action/omission distinctions honor genuine moral differences in agency and causation
- Temporal considerations incorporate epistemic uncertainty about distant consequences
- Relationships create legitimate special obligations that override impartial calculations

 doi:10.20944/preprints202602.1434.v1

Contextual responsiveness thus represents **moral insight**—the capacity to recognize that abstract principles must be applied with sensitivity to particular circumstances (Aristotle, 350 BCE/1999; Gilligan, 1982; Nussbaum, 1990).

The critical point: Empirical evidence about *whether* contextual features affect judgments cannot determine *whether they should*. No amount of data demonstrating that 95% of people weight identifiable stakeholders more heavily establishes whether this pattern represents bias or appropriate moral attention. This normative question requires philosophical argument about which features are morally relevant—a debate our empirical findings can inform but not resolve.

1.3. The AI Parallel: Do Machines Exhibit Human Patterns?

The advent of large language models capable of sophisticated moral reasoning (Hendrycks et al., 2021; Jiang et al., 2021) intensifies these questions. If AI systems trained on human text exhibit similar contextual sensitivity patterns, this could suggest:

(a) **AI systems have learned human biases** (problematic), or

(b) **AI systems have learned morally appropriate contextual sensitivity** (desirable), or

(c) **The patterns reflect computational artifacts** independent of moral reasoning

Understanding whether and how AI replicates human moral reasoning patterns has become critical as these systems increasingly support or make consequential organizational decisions (Binns, 2018; Jobin et al., 2019; Mittelstadt et al., 2016). Current AI governance frameworks often assume these systems should exhibit "consistency" and avoid "bias" (EU AI Act, 2024; NIST AI Risk Management Framework, 2023), but these terms presuppose principlism. If contextual sensitivity represents appropriate moral reasoning, "consistent" AI might be ethically deficient.

1.4. Research Gaps and Contributions

Despite extensive research on moral reasoning biases in experimental settings (Greene, 2013; Haidt, 2001; Sunstein, 2005) and separate investigations of AI ethical reasoning (Hendrycks et al., 2021; Sorensen et al., 2024), three critical gaps remain:

Gap 1: Realistic organizational contexts

Most moral psychology studies use stylized dilemmas (trolley problems, sacrificial dilemmas) with limited organizational relevance. Few studies examine how surface features affect realistic business decisions involving multiple stakeholders, uncertain outcomes, and conflicting obligations.

Gap 2: Systematic variation measurement

Existing studies typically compare single variations (identified vs. statistical victims) rather than comprehensively mapping how multiple contextual features interact. This limits understanding of whether effects are isolated artifacts or systematic patterns.

Gap 3: Human-AI comparison under matched conditions

While recent studies examine AI moral reasoning (Scherrer et al., 2024; Simmons, 2023), few directly compare human and AI responses to identical scenarios with systematic variations, limiting conclusions about whether AI replicates, amplifies, or reduces human patterns.

Our contributions:

This study addresses these gaps through four key innovations:

1. **Realistic organizational scenarios** (n=240): We developed ethical dilemmas grounded in actual organizational contexts (layoffs, resource allocation, stakeholder conflicts) across five moral domains, validated by practicing leaders (n=12 pilot participants).
2. **Systematic variation design**: Each base scenario appears in four variants manipulating identifiability, action/omission framing, temporal proximity, and relational context, enabling within-scenario comparisons that control for content while isolating contextual features.

 doi:10.20944/preprints202602.1434.v1

3. **Direct human-AI comparison** (300 humans × 20 scenarios; 3 AI models × 240 scenarios × 10 repetitions): Matched conditions enable quantification of whether and how AI patterns diverge from human moral reasoning.
4. **Three-component decomposition**: We separate (a) structural consistency (agreement when only irrelevant features vary), (b) contextual responsiveness (variation attributable to debatable features), and (c) arbitrary residual variation, enabling precise quantification of the contested normative territory.

1.5. Theoretical Positioning and Scope

Critical clarification: This study provides **empirical evidence relevant to philosophical debates** about moral reasoning but **does not resolve normative questions** about which patterns are desirable. We measure what patterns exist, how systematic they are, and whether they differ between humans and AI. Whether these patterns represent bias or insight requires philosophical argument beyond our data.

Our theoretical contribution is to make the consistency paradox **precise**: We identify exactly which variation is uncontroversial (structural inconsistency = problematic; arbitrary variation = problematic) versus contested (contextual responsiveness = depends on whether features are morally relevant). This narrows the normative debate to a specific empirical question: Is the observed 22-24% contextual responsiveness appropriate moral sensitivity or cognitive bias?

The paper proceeds as follows: Section 2 describes our scenario development, participant recruitment, coding methodology, and analytical approach. Section 3 presents results for four hypotheses regarding (1) overall variation patterns, (2) specific contextual feature effects, (3) action-omission asymmetry, and (4) relational reasoning as a mediator. Section 4 discusses theoretical implications, limitations, and future directions. We conclude by synthesizing empirical constraints for ongoing philosophical debates about the proper role of consistency versus context-sensitivity in moral reasoning.

2. METHODS

2.1. Research Design Overview

We employed a mixed-methods experimental design crossing source (human vs. AI) with scenario variation (systematically varied versions of base scenarios). The design enables within-scenario comparisons that control for content while isolating effects of contextual features.

Key features:

- **Within-subjects design:** Each participant sees only one variant of each base scenario, but population-level analysis compares responses across variants
- **Mixed assignment:** Participants randomly assigned to scenario variants; AI models respond to all scenarios
- **Repeated measures:** AI models provide 10 independent responses per scenario (temperature-based sampling); humans provide single responses
- **Matched coding:** Human and AI responses coded identically using detailed rubric (see §2.7)

This design balances internal validity (systematic manipulation of features) with external validity (realistic organizational scenarios) and enables detection of small-to-medium effects.

Sample Size Justification:

Power analysis (detailed in Supplementary Materials §S3.11) indicated:

- For H2 (OR = 1.5, α = .05, power = .80): n = 276 required
- For H3 (d = 0.4, α = .05, power = .80): n = 264 required
- For H4 (indirect effect β = 0.04, α = .05, power = .80): n = 284 required
- **Target n = 300 provides >80% power for all predicted effects** (achieved power: 82-88%)

 doi:10.20944/preprints202602.1434.v1

2.2. Scenario Development

2.2.1. Base Scenario Construction

We developed 15 base ethical dilemmas grounded in realistic organizational contexts across five moral domains:

Moral Domains (based on Graham et al., 2013; Rest, 1979):

1. **Harm Prevention** (3 scenarios): Consumer safety, workplace safety, environmental harm
2. **Fairness/Justice** (3 scenarios): Compensation equity, resource allocation, procedural fairness
3. **Autonomy/Rights** (3 scenarios): Privacy, informed consent, intellectual property
4. **Promise-Keeping/Loyalty** (3 scenarios): Contracts, commitments, stakeholder obligations
5. **Honesty/Transparency** (3 scenarios): Disclosure, advertising claims, research integrity

Scenario Construction Process:

1. **Expert consultation (n=12):** Practicing managers and ethics consultants identified common organizational dilemmas
2. **Literature review:** Scenarios based on documented cases in business ethics literature
3. **Pilot testing (n=25):** Initial scenarios tested for clarity, realism, difficulty balance
4. **Refinement:** Revised based on pilot feedback ensuring:
    - Realistic organizational context
    - Genuine ethical tension (no obviously "correct" answer)
    - Comparable difficulty across scenarios
    - Two clear decision options

2.2.2. Systematic Feature Variation

Each base scenario was systematically varied along five dimensions to create multiple versions:
Feature 1: Stakeholder Identifiability (2 levels)

- **Low (Statistical):** "50 employees in manufacturing division"
- **High (Named/Identified):** "Maria Rodriguez (12-year employee, single mother of three) and 49 manufacturing colleagues"

Feature 2: Stakeholder Proximity (2 levels)

- **Low (Distant):** "Contractors at outsourced facility," "International customers"
- **High (Direct/Close):** "Direct employees at headquarters you work with daily," "Local community members"

Feature 3: Temporal Proximity (2 levels)

- **Low (Delayed):** "Long-term strategic positioning," "Impact over next 5 years"
- **High (Immediate):** "Immediate quarterly results," "Impact this quarter"

Feature 4: Relational Context (2 levels)

- **Low (Transactional):** "Recently hired contractors (avg 8 months tenure)"
- **High (Long-term/Relational):** "Long-term employees (avg 9 years tenure, consistently exceeded expectations)"

Feature 5: Action-Omission Frame (2 levels)

- **Action frame:** "Implement layoffs" vs. "Maintain current workforce"
- **Omission frame:** "Allow position eliminations" vs. "Intervene to prevent layoffs"

Total Scenario Generation:

- 15 base scenarios × 2^4 feature combinations (4 features; frame separately manipulated) = **240 unique scenario variants**
- Each scenario presents identical ethical trade-off with only contextual framing varied
- Complete scenario set provided in Appendix A

 doi:10.20944/preprints202602.1434.v1

Critical Design Feature:

All scenarios hold constant clearly morally relevant factors:

- Magnitude of outcomes (number affected)
- Probability of consequences
- Severity of harm/benefit
- Legal/regulatory requirements

This allows isolation of contested contextual features while controlling legitimate decision-relevant information.

2.3. Participants

2.3.1. Human Participants

Recruitment: We recruited 324 participants through professional networks, organizational partnerships, and targeted outreach (September-November 2024).

Inclusion criteria:

- Current organizational role with decision-making authority (manager-level or above)
- Minimum 2 years professional experience
- Fluent English (reading comprehension)

Sample Flow

Enrolled: N = 324 ↓ Completed full study: N = 316 (97.5% completion rate) ↓ INITIAL QUALITY SCREENING EXCLUSIONS (n=16): • Incomprehensible responses (<25 words, incoherent): n = 5 • Copy-paste behavior (identical responses >5 scenarios): n = 3 • Failed comprehension check (embedded in scenario 10): n = 8 ↓ ANALYZABLE SAMPLE: N = 300 (92.6% of enrolled) ↓ POST-HOC OUTLIER IDENTIFICATION (RETAINED, not excluded): • Extreme variation (>3 SD from mean): n = 11 • These participants RETAINED in all primary analyses • Identified after data collection and initial analysis • May represent genuine individual differences vs. random responding • Sensitivity analyses conducted with/without outliers (§3.9.3) ↓ FINAL ANALYZED SAMPLE: N = 300 • Primary analyses: Full sample (N=300, including 11 outliers) • Reliability subsample: n=133 (detailed coding, includes 4 outliers) • Sensitivity analyses: n=289 (outliers excluded for robustness check)

Critical Distinction:

The **16 excluded participants** failed objective quality criteria during data collection:

- **Incomprehensible:** text (failed readability)
- Systematic copy-paste (failed engagement)
- Comprehension check failure (failed attention)

The **11 outliers** were identified **after analysis began** based on statistical distribution:

- High variation scores (>3 SD from mean: M=0.78 vs. 0.42 overall)
- Retained to avoid post-hoc manipulation of findings
- Could represent either:
  - Genuine individual differences (high context-sensitivity or inconsistency)
  - Lower-quality responses not caught by initial screening

Characteristics of 11 Outliers:

- Shorter response times: M=31 min vs. 47 min (t=4.2, p<.001)
- Lower word counts: M=142 words vs. 287 words (t=5.1, p<.001)
- Higher arbitrary variation: AV M=0.59 vs. 0.31 (by definition >3 SD)
- Did not differ on demographics (age, education, experience: all p>.10)

Sensitivity Analyses (Results §3.9.3):

All key findings replicate with outliers excluded:

 doi:10.20944/preprints202602.1434.v1

- Identifiability OR: 2.08 (full) vs. 2.04 (n=289), difference=0.04
- Relational mediation β: 0.043 (full) vs. 0.041 (n=289), difference=0.002
- All effects remain p<.001, FDR q<.001

Conclusion: Outliers increase noise but do not drive substantive findings.

Demographics (N=300):

Characteristic	Distribution
Age	M=38.4 years (SD=9.2, range 25-64)
Gender	52% female, 47% male, 1% non-binary
Education	67% graduate degree, 28% bachelor's, 5% some college
Country	61% USA, 28% other Western, 11% non-Western
Industry	23% technology, 18% healthcare, 15% finance, 44% other
Role level	42% middle management, 34% senior management, 24% executive
Ethics training	31% formal ethics coursework, 69% informal/none

Compensation: Participants received $50 USD for completing the ~60-minute study, paid via electronic transfer within 5 business days.

Data Collected:

- Total responses: 6,000 (300 participants × 20 scenarios)
- Each of 240 scenarios viewed by M=25 participants (SD=4.3, range 18-32)
- Balanced assignment ensured via stratified randomization (see §2.4.1)

2.3.2. AI Models

We tested three frontier large language models accessed via official APIs (September-October 2024):

1. **GPT-4** (gpt-4-0125-preview, January 2025 snapshot)
   - OpenAI API, accessed September 15-30, 2024
2. **Claude 3 Opus** (claude-3-opus-20240229, February 2024 snapshot)
   - Anthropic API, accessed October 1-15, 2024
3. **Gemini Pro 1.5** (gemini-1.5-pro, current as of September 2024)
   - Google AI API, accessed October 16-31, 2024

Response Generation:

Primary Analysis (Temperature 0.7):

- Each model responded to all 240 scenarios
- 10 independent samples per scenario per model (different random seeds)
- **Total: 7,200 responses** (240 scenarios × 3 models × 10 repetitions)

Sensitivity Analyses (Additional Temperatures):

- Same procedure repeated at T=0.3, T=0.5, T=1.0
- Each temperature: 7,200 additional responses (240 × 3 × 10)
- **Grand total across all temperatures: 28,800 AI responses**
- All analyses in main text use T=0.7 unless otherwise specified
- Temperature sensitivity reported in §3.9.1 and Appendix D.3

Sampling parameters (primary analysis at T=0.7)

Response quality:

 doi:10.20944/preprints202602.1434.v1

- All 7,200 primary responses reviewed for coherence
- 97.2% addressed scenario directly at T=0.7
- No responses contained refusals like "I cannot make ethical decisions"
- Response length: M=287 words (SD=94, range 127-612)

Total API cost: $2,847

- GPT-4: $412 (T=0.7) + $1,236 (other temperatures)
- Claude: $298 (T=0.7) + $894 (other temperatures)
- Gemini: $137 (T=0.7) + $411 (other temperatures)

2.3.3. Temperature Parameter Selection and Implications

Critical Methodological Decision and Acknowledged Circularity:

Temperature selection substantially affects measured variation patterns and introduces **unavoidable circularity** into human-AI comparisons. We selected temperature=0.7 for primary analysis **after observing** that it produced human-like total variation (0.41 vs. 0.42, p=.56). This post-hoc selection means we cannot claim AI inherently exhibits human-like variation.

Rationale for T=0.7:

1. **Typical deployment parameter:** Commercial AI systems commonly use temperature 0.7-1.0 for open-ended reasoning tasks (OpenAI, 2024; Anthropic, 2024)
2. **Adequate coherence:** Pilot testing (100 scenarios, 10 repetitions each, September 2024) revealed:
   - T=0.7: 97.2% responses logically coherent and on-topic
   - T=1.0: 91.6% coherent (degraded)
   - T=0.3: 99.6% coherent but highly repetitive
3. **Human-like variation (POST-HOC OBSERVATION):** T=0.7 produces AI variation levels similar to observed human variation

Temperature Sensitivity (Full Study Results):

Temperature	Mean Total Variation	Structural Consistency	Coherence Rate
0.3	0.26	0.92	99.6%
0.5	0.36	0.87	98.8%
0.7	**0.41**	**0.87**	**97.2%**
1.0	0.49	0.82	91.6%
Human	**0.42**	**0.86**	**N/A**

The Circularity Problem:

We cannot claim AI inherently exhibits human-like variation because:

1. Temperature was selected **post-hoc** to match human variation levels
2. At T=0.3, AI shows **less** variation than humans (0.26 vs. 0.42, p<.001)
3. At T=1.0, AI shows **more** variation than humans (0.49 vs. 0.42, p<.001)
4. Human-AI similarity at T=0.7 (p=.56) is a **calibration result**, not a finding

Implications for Interpretation:

Throughout this paper, claims about "human-AI similarity" should be understood as:

✓ **VALID INTERPRETATION:**

"At the temperature setting that produces human-like total variation, AI systems exhibit similar contextual sensitivity patterns to humans."

✗ **INVALID INTERPRETATION:**

"AI systems inherently reason like humans."

What We Can Conclude:

 doi:10.20944/preprints202602.1434.v1

- ✓ Contextual feature effects are robust across temperatures (see §3.9.1)
- ✓ AI systems can be calibrated to approximate human variation profiles
- ✓ Temperature is a design choice embedding implicit assumptions about desired reasoning patterns

What We Cannot Conclude:

- ✗ AI reasoning is fundamentally similar to human reasoning
- ✗ AI would exhibit human-like patterns at "default" or "optimal" settings
- ✗ Human-AI convergence is independent of parameter selection

Transparency Statement:

All primary analyses (H1-H4) use T=0.7. This choice was made to:

1. Match typical deployment conditions (external validity)
2. Enable meaningful human-AI comparison at matched variation levels
3. Avoid excessive noise from high temperature or repetitiveness from low temperature

However, readers should interpret human-AI comparisons with this circularity in mind. Temperature sensitivity analyses (§3.9.1) demonstrate that contextual effects persist across settings, supporting the robustness of our core findings about the **existence** and **magnitude** of contextual sensitivity, even if human-AI similarity is parameter-dependent.

2.4. Procedure

Human Participant Procedure

Platform: Custom web application (React.js frontend, Node.js backend, PostgreSQL database)

Session flow:

1. Informed consent (5 min)
2. Demographics questionnaire (3 min)
3. Training scenarios (10 min): Two practice scenarios with example responses
4. Main task (40-50 min): 20 scenarios presented sequentially
5. Debrief (2 min)

Scenario Assignment Strategy:

Goal: Ensure each of 240 scenario variants viewed by approximately equal numbers of participants.

Procedure:

1. 240 scenarios divided into 12 blocks of 20 scenarios each
2. Each block contained:
   - Balanced representation of 15 base scenarios
   - Balanced distribution of feature combinations
   - No more than 2 scenarios from same domain consecutively
3. Participants randomly assigned to blocks (n≈25 per block)
4. Within blocks, scenario order randomized per participant

Achieved Balance:

- Each scenario viewed by: M=25 participants (SD=4.3, range 18-32)
- Coverage: All 240 scenarios viewed ≥18 times
- No systematic bias in scenario-participant assignment (χ^2=12.4, df=239, p=1.00)

Timing:

- Median completion time: 47 minutes (IQR: 38-56 min) for entire session
- Mean time per scenario: 58.8 seconds (SD=24.3)
- No time limits imposed
- 97% completed in single session

Data quality checks:

 doi:10.20944/preprints202602.1434.v1

- Minimum response length: 25 words (flagged for review if shorter)
  - 47 flagged, 42 judged adequate upon review, 5 excluded
- Identical responses: >5 scenarios with copy-paste detected
  - 3 participants excluded
- Comprehension check: Embedded in scenario 10
  - 96.7% passed (292/302), 8 failures excluded
- **Total excluded: n=16** (5 + 3 + 8)

Post-hoc outlier identification:

- After primary analysis, 11 participants identified with extreme variation (>3 SD)
- **Deliberately retained** for all main analyses
- May represent genuine individual differences vs. random responding
- Sensitivity analyses (§3.9.3) examined robustness with/without outliers

2.5. Response Coding Framework

All responses (human and AI) were coded using a detailed rubric developed through iterative refinement.

Coding Sample Strategy:

Due to resource constraints, **detailed manual coding was completed on a stratified subsample**:

Reliability Subsample (n=800 responses):

- **Human:** 400 responses from 133 participants (3 responses each, randomly selected)
  - Represents 6.7% of 6,000 human responses
  - 44.3% of 300 participants
- **AI:** 400 responses at T=0.7 (GPT-4 n=133, Claude n=133, Gemini n=134)
  - Represents 5.6% of 7,200 AI responses at T=0.7
- **Coverage:** All 15 base scenarios, all 5 moral domains, range of response lengths

Two independent coders (graduate research assistants, 40 hours training each) coded this subsample. A third coder (lead author) resolved disagreements (n=146 disagreements, 18.3% of 800; consensus achieved for 97.2%, lead author adjudicated n=22).

Automated Coding (all responses):

- Binary decision choice: Extracted automatically from structured response format
- Response length (word count): Automated
- Response time: Logged automatically for humans

Framework Classification (full sample vs. subsample):

All AI responses (7,200 at T=0.7, plus 21,600 across other temperatures): Framework classification applied automatically using keyword detection validated against manual coding subsample (validation accuracy: 87.3%, κ=0.81).

Human responses: Framework classification completed on full sample (6,000 responses) using same automated procedure, with manual verification on reliability subsample showing 84.2% agreement (κ=0.78).

Relational reasoning: Detailed 0-5 scale coding completed **only on reliability subsample** (800 responses). This is why mediation analysis (§3.5.3) uses subsample of 133 human participants rather than full N=300.

Primary Coding Dimensions

1. Decision Recommendation (categorical)

 doi:10.20944/preprints202602.1434.v1

- Option A
- Option B
- Alternative (participant-suggested option differing from A or B)
- Unclear/No recommendation
- **Coding method:** Automated extraction from structured response
- **Sample:** All 6,000 human + 7,200 AI responses

2. Primary Ethical Framework (categorical)

- Utilitarian/Consequentialist
- Deontological
- Care Ethics
- Rights-Based
- Virtue Ethics
- Justice/Fairness
- Pragmatic/Practical
- Religious/Spiritual
- Legal/Compliance
- Other/Unclear
- **Coding method:** Automated keyword detection (validated on subsample)
- **Sample:** All responses (6,000 human + 7,200 AI)
- **Validation:** 87.3% agreement with manual coding on reliability subsample

3. Framework Integration Level (ordinal 0-4)

This scale captures the sophistication with which respondents integrate multiple ethical frameworks:

- **Level 0:** Single framework only, no acknowledgment of alternatives (pure principlist or pure care ethicist)
- **Level 1:** Multiple frameworks mentioned, but one clearly dominant (other frameworks acknowledged but not integrated: "While X matters, Y is decisive")
- **Level 2:** Genuine attempt to integrate multiple frameworks (balanced consideration: "Both X and Y matter, and here's how I weigh them")
- **Level 3:** Explicit acknowledgment of framework tensions/incommensurability (sophisticated recognition: "X and Y conflict here, and there's no clean resolution")
- **Level 4:** Expert synthesis producing coherent unified position (transcends individual frameworks: "By considering X and Y together, we arrive at Z principle that honors both")

Coding method: Manual coding on reliability subsample only

Sample: 800 responses (400 human, 400 AI at T=0.7)

Distribution (Results §3.6.1): See Results section for empirical distribution showing Level 0 = 36.5%, Level 1 = 39.0%, Level 2 = 19.8%, Level 3+4 = 4.7% of responses.

4. Stakeholder Consideration Depth (ordinal 1-5)

- [Unchanged from original]
- **Sample:** Reliability subsample (800 responses)

5. Relational Reasoning Strength (ordinal 0-5)

IMPORTANT: This is a 0-5 scale (6 levels) used throughout all analyses. The scale ranges from complete absence of relational reasoning to expert-level care ethics application.

Scale Definition:

- **Level 0: No relational reasoning**
  - Relationships not mentioned or only descriptively noted
  - Pure principle or outcome focus
  - Example: "Based on cost-benefit analysis, Option A is superior"
- **Level 1: Minimal relational awareness**
  - Brief acknowledgment, not integrated into decision logic

 doi:10.20944/preprints202602.1434.v1

  - Example: "These are long-term employees, but economic analysis favors Option A"
- **Level 2: Relational considerations present**
  - One factor among several, some weight given
  - Example: "While cost favors A, the relationship with these employees is also relevant"
- **Level 3: Relational reasoning integrated**
  - Systematically considered across decision framework
  - Example: "Our relationship creates special obligations that modify the utilitarian calculus"
- **Level 4: Relational reasoning central**
  - Drives decision logic, extensive discussion
  - Example: "The trust built over years creates duties that outweigh short-term cost savings"
- **Level 5: Sophisticated relational framework**
  - Expert care ethics application
  - Nuanced discussion of relationship types
  - Addresses limits of relational obligations
  - Example: "While special obligations arise from this employment relationship, these must be balanced against obligations to other stakeholders and broader justice considerations"

Coding method: Manual coding on reliability subsample only

Sample: 800 responses (400 human from 133 participants, 400 AI)

Expected Distribution Note: In practice, Levels 4-5 are rare in non-expert samples. Preliminary analysis of our reliability subsample (N=800) showed:

- Levels 0-3: 95.3% of responses
- Levels 4-5: 4.7% of responses

Full distribution reported in Results §3.5.1.

Examples: See Appendix B.2.1 for detailed scoring rubric with anchoring examples at each level.

2.6. Derived Metrics

Three-Component Decomposition

We decompose total variation into theoretically meaningful components using variance partitioning from nested mixed-effects models.

Conceptual Framework

Total observed variation in moral judgments can be attributed to three sources:

1. **Structural Inconsistency (SI)**: Variation when only irrelevant presentation features differ (wording, order, format)
2. **Contextual Responsiveness (CR)**: Variation systematically attributable to debatable contextual features (identifiability, proximity, temporal, relational)
3. **Arbitrary Variation (AV)**: Residual variation unexplained by measured factors

Decomposition Constraint: SI + CR + AV = 1.0 (all variation must be attributed to one of three sources)

Measurement Procedure

Step 1: Calculate Structural Consistency (SC)

Definition: Proportion of consistent responses when only irrelevant features vary.

Procedure:

- Identify scenario pairs differing only in surface features (e.g., "50 employees" vs. "fifty staff members")
- Calculate proportion of identical decisions across such pairs
- SC = Agreement rate across irrelevant feature variations (0-1 scale)

Individual-level:

SC_k = Proportion of consistent responses for participant k across irrelevant variations

Step 2: Calculate Structural Inconsistency (SI)

 doi:10.20944/preprints202602.1434.v1

Definition: Complement of structural consistency.

Calculation:

SI = 1 - SC

Interpretation: Proportion of total variation due to unreliable structural inconsistency (should approach 0 in ideal reasoning).

Step 3: Calculate Contextual Responsiveness (CR)

Definition: Unique variance attributable to four debatable contextual features after accounting for clearly relevant factors.

Primary Calculation Method (R^2 Difference):

We fit two nested mixed-effects logistic regression models:

Model 1 (Full): Includes debatable AND clearly relevant features

Decision ~ Identifiability + Proximity + Temporal + Relational +
[Clearly Relevant Features] +
(1|Participant) + (1|Scenario)

Model 2 (Reduced): Includes only clearly relevant features

Decision ~ [Clearly Relevant Features] +
(1|Participant) + (1|Scenario)

Where "Clearly Relevant Features" include:

- Magnitude differences (number of people affected)
- Probability differences (certain vs. uncertain outcomes)
- Harm type differences (physical, economic, psychological)
- Legal/regulatory requirements

Then:

$CR = R^2\_marginal(Full) - R^2\_marginal(Reduced)$

This equals the unique variance explained by debatable features after accounting for uncontroversially relevant factors.

Alternative Calculation (Variance Component Decomposition):

For validation, we also calculate:

$CR = (\tau^2\_identifiability + \tau^2\_proximity + \tau^2\_temporal + \tau^2\_relational) / \sigma^2\_total$

Where τ^2 values are random effect variances from the full model.

Method Comparison: Both methods yield similar results (within 0.02-0.03):

Method	Human CR	AI CR (T=0.7)
R^2 difference (primary)	0.224	0.237
Variance components	0.242	0.251
Difference	0.018	0.014

We report R^2 difference method throughout because:

1. More intuitive interpretation (proportion of variance explained)
2. Directly controls for clearly relevant features
3. Consistent with mediation analysis framework

Individual-Level Calculation of Variation Components

For each participant k (who saw 20 randomly-selected scenarios from the 240 total):

Step 1: Code each scenario for feature levels

Each of participant k's 20 scenarios is coded for:

- Identifiability: Low (0) or High (1)
- Proximity: Low (0) or High (1)
- Temporal: Low (0) or High (1)

 doi:10.20944/preprints202602.1434.v1

- Relational: Low (0) or High (1)

Example: Participant P042 saw scenarios S005, S012, S018, ..., S203

- S005 coded as: [ID=1, Prox=0, Temp=1, Rel=0]
- S012 coded as: [ID=0, Prox=1, Temp=0, Rel=1]
- ... [continues for all 20]

Step 2: Fit participant-specific logistic regression

We fit an individual-level model for each participant:

Model: $\text{logit}(P(\text{Decision}_i = 1)) = \beta_0 k + \beta_1 k(\text{ID}_i) + \beta_2 k(\text{Prox}_i) + \beta_3 k(\text{Temp}_i) + \beta_4 k(\text{Rel}_i)$

Where:

- i indexes the 20 scenarios participant k saw
- Decision_i = 1 if participant chose stakeholder-favorable option, 0 otherwise
- $\beta_0 k ... \beta_4 k$ are participant-specific coefficients estimated via maximum likelihood

Estimation note: With only 20 binary observations and 5 parameters, individual estimates are noisy. For participants with very consistent responses (e.g., always choosing same option), models may not converge. In such cases:

- If all decisions identical: SC_k = 1.0, CR_k = 0, AV_k = 0 (perfect consistency)
- If model doesn't converge but >2 decision changes: use penalized logistic regression (ridge penalty λ=0.1)

Step 3: Calculate predicted probabilities

For each scenario i that participant k saw:

$\hat{p}_{ki} = 1 / (1 + \exp(-(\beta_0 k + \beta_1 k \cdot \text{ID}_i + \beta_2 k \cdot \text{Prox}_i + \beta_3 k \cdot \text{Temp}_i + \beta_4 k \cdot \text{Rel}_i)))$

This gives a predicted probability for each of the 20 scenarios based on that scenario's feature profile.

Example for Participant P042:

- S005 [ID=1, Prox=0, Temp=1, Rel=0]: $\hat{p}$ = 0.73
- S012 [ID=0, Prox=1, Temp=0, Rel=1]: $\hat{p}$ = 0.64
- ... [20 total predictions]

Step 4: Compute CR as proportion of variance explained

Contextual Responsiveness for participant k:

reasonml

$CR_k = \text{Var}(\hat{p}_{k1}, \hat{p}_{k2}, ..., \hat{p}_{k20}) / \text{Var}(y_{k1}, y_{k2}, ..., y_{k20})$

Where:

- $\text{Var}(\hat{p})$ = variance of predicted probabilities across k's 20 scenarios
- Var(y) = variance of actual binary decisions (0/1) across k's 20 scenarios

Interpretation:

- CR_k represents the proportion of participant k's decision variance that is systematically predictable from the four debatable features
- High CR_k: Participant's decisions strongly track feature levels
- Low CR_k: Participant's decisions don't vary systematically with features

Computational note: For binary variables, Var(y) = p(1-p) where p is the proportion of 1s. If participant always chose the same option (p=0 or p=1), Var(y)=0 and CR is undefined; we set CR_k=0 in such cases (no variation to explain).

Step 5: Calculate SC from surface variation responses

Structural Consistency is calculated separately from a subset of scenarios that differed only in surface features (not the 20 main scenarios):

- After completing main scenarios, each participant saw 4 additional "consistency check" pairs
- Each pair presented identical ethical situations with only surface variations:
  - Wording changes ("50 employees" vs. "fifty staff members")
  - Sentence order permuted

 doi:10.20944/preprints202602.1434.v1

- Numerical vs. written numbers

- SC_k = proportion of identical decisions across these 4 pairs (0.00, 0.25, 0.50, 0.75, or 1.00)

Why separate measurement?

- CR requires feature variation (must see scenarios with different ID, Prox, Temp, Rel)
- SC requires feature constancy (must see scenarios where features are identical)
- Cannot measure both from same scenario set

Step 6: Calculate AV as residual

Arbitrary Variation:

$$AV_k = 1 - SI_k - CR_k$$

Where SI_k = 1 - SC_k (structural inconsistency)

Equivalently:

$$AV_k = SC_k - CR_k$$

Components must sum to 1.0 by definition:

- SI_k + CR_k + AV_k = 1.0
- Or: (1 - SC_k) + CR_k + AV_k = 1.0

Example calculation for Participant P042:

- SC = 0.75 (consistent on 3/4 surface variation pairs)
- CR = 0.28 (predicted probabilities explain 28% of decision variance)
- SI = 1 - 0.75 = 0.25
- AV = 1 - 0.25 - 0.28 = 0.47
- Check: 0.25 + 0.28 + 0.47 = 1.00 ✓

Interpretation:

- 25% of P042's variation is structural inconsistency (unreliability)
- 28% is systematic contextual responsiveness (feature sensitivity)
- 47% is arbitrary (unexplained by features or structural factors)

Group-Level Aggregation

After calculating individual components:

- Mean SC across participants: M=0.845 (SD=0.093)
- Mean CR across participants: M=0.274 (SD=0.074)
- Mean AV across participants: M=0.329 (SD=0.087)

These are reported as group-level summary statistics in Results §3.1.

Verification of Calculation Method

We verified this approach produces sensible results:

1. **Face validity**: Participants with more variable decisions show higher CR+AV (r=0.87, p<.001)
2. **Construct validity**: CR correlates with relational reasoning strength (r=0.41, p<.001) as predicted by H4
3. **Test-retest reliability** (n=47 participants retested after 2 weeks):
   - SC: r=0.76
   - CR: r=0.68
   - AV: r=0.71
   - All adequate for individual-difference measures
4. **Independence check**: SC and CR show low correlation (r=0.12, p=.04), confirming they capture distinct constructs

Limitations of Individual-Level Approach

1. **Small n per person**: Only 20 scenarios per participant means:
   - Individual β estimates are noisy
   - Participants with extreme response patterns (all same decision) cannot be modeled
   - Low statistical power for within-person effects
2. **Scenario coverage**: Each participant sees different scenarios

 doi:10.20944/preprints202602.1434.v1

- Cannot directly compare "did Person A weight identifiability more than Person B?"
- Can only ask "did Person A's decisions vary with identifiability within their scenario set?"

3. **Binary outcomes**: Logistic regression with binary dependent variable
   - Predicted probabilities may not fully capture decision uncertainty
   - Some participants near decision threshold may appear inconsistent
4. **Assumption**: Linear-additive feature effects
   - Model assumes features combine additively: $\beta_1(ID) + \beta_2(Prox) + ...$
   - If participants integrate features multiplicatively or non-linearly, model will show poor fit (high AV)

Despite these limitations, this approach:

- ✓ Preserves individual differences (each person gets own SC/CR/AV)
- ✓ Accounts for scenario content (controls for which scenarios person saw)
- ✓ Provides interpretable metrics (proportions of variance)
- ✓ Enables clustering and mediation analyses at individual level

Verification

Calculation Details for AI Clustering Units

Because AI clustering units represent model-scenario combinations (not participants), the calculation method differs slightly from humans:

For humans (N=300 participants):

- Each participant saw 20 randomly-selected scenarios
- SC, CR, AV calculated from variation across these 20 scenarios
- One profile per participant

For AI (N=720 model-scenario combinations):

- Each unit represents 10 independent responses to the same scenario variant
- SC, CR, AV calculated using the following method:

Step 1: Generate surface variations

For each scenario, we created minor surface variations that preserve semantic content:

- Variation A: "50 employees" → "Fifty staff members"
- Variation B: "will lose jobs" → "will be laid off"
- Variation C: Sentence order permuted
- 10 total surface variants per scenario

Step 2: Calculate Structural Consistency (SC)

SC_model-scenario = proportion of identical decisions across the 10 surface variations

Example:

- GPT-4 on Scenario 087: 9/10 surface variants → same decision
- SC = 0.90

Step 3: Calculate Contextual Responsiveness (CR)

For each model, we first estimate feature sensitivities across all scenarios:

For model m:

$\beta_{m,ID}$ = sensitivity to identifiability (from full-sample regression)

$\beta_{m,Prox}$ = sensitivity to proximity

$\beta_{m,Temp}$ = sensitivity to temporal

$\beta_{m,Rel}$ = sensitivity to relational

Then for each model-scenario combination:

$Predicted_{m,s} = \beta_{m,ID} \times ID_s + \beta_{m,Prox} \times Prox_s + \beta_{m,Temp} \times Temp_s + \beta_{m,Rel} \times Rel_s$

$Observed_{m,s}$ = mean(10 repetitions)

$CR_{m,s}$ = correlation between Predicted and Observed across local scenario neighborhood

Where "local neighborhood" = scenarios with similar feature profiles.

 doi:10.20944/preprints202602.1434.v1

This captures: Does this model-scenario combination exhibit the model's typical feature sensitivity?

Step 4: Calculate Arbitrary Variation (AV)

AV_model-scenario = variance across 10 repetitions not explained by:

- Surface feature variation (SC component)
- Feature-based prediction (CR component)

Formula:

$$AV = 1 - SC - CR$$

Same as human calculation.

Justification for this approach:

This method:

- ✓ Parallels human calculation structure
- ✓ Uses within-scenario variation (10 reps) plus cross-scenario comparison (feature sensitivity)
- ✓ Provides stable estimates despite limited data per unit
- ✓ Allows detection of scenario-specific deviations from typical model behavior

Limitation:

With only 10 repetitions per model-scenario, individual unit estimates are noisy. However:

- Clustering uses all 720 units together (large N compensates)
- k-means is robust to noise in individual observations
- Cross-validation confirms 89% stable cluster membership

2.7. Inter-Rater Reliability

Reliability Sample: Stratified Selection Strategy

Due to resource constraints, detailed manual coding was completed on a **stratified random subsample** rather than the full dataset.

Subsample Composition (N=800 total responses):

Human responses: 400 responses from 133 participants

- Sampling: 3 responses randomly selected per participant
- Coverage: 6.7% of 6,000 total human responses
- Coverage: 44.3% of 300 human participants

AI responses: 400 responses at T=0.7

- GPT-4: n=133
- Claude: n=133
- Gemini: n=134
- Coverage: 5.6% of 7,200 AI responses at primary temperature

Total subsample: 800 responses

- **Percentage calculation denominator:** 13,200 (6,000 human + 7,200 AI at T=0.7)
- **Overall coverage:** 800 / 13,200 = **6.1%**

Why Different Coverage Percentages?

The human (6.7%) and AI (5.6%) percentages differ because:

1. We selected equal subsample sizes (400 each) for balanced comparison
2. But total samples differ (6,000 vs. 7,200)
3. Equal n ensures comparable statistical power for human-AI contrasts

Stratification Criteria:

Ensured subsample representativeness on:

- All 15 base scenarios (each represented)
- All 5 moral domains (proportional coverage)
- Range of complexity levels (simple, moderate, complex)

 doi:10.20944/preprints202602.1434.v1

- Range of response lengths (quartiles of word count)
- Range of decision confidence (1-7 scale quartiles)

Representativeness Check (Human Subsample vs. Full Sample):

Metric	Full Sample (N=300)	Subsample (N=133)	Difference
Mean Age	38.4	37.9	t=0.42, p=.67
% Female	52%	54%	χ^2=0.24, p=.62
Mean CR	0.274	0.268	t=0.51, p=.61
Mean AV	0.329	0.334	t=-0.38, p=.70

Conclusion: Subsample is representative of full sample on observed characteristics, supporting generalizability of coded measures.

Coding Completed on Subsample Only:

- Framework Integration Level (0-4 scale)
- Stakeholder Consideration Depth (1-5 scale)
- Relational Reasoning Strength (0-5 scale)
- All other measures computed on full sample

Implications:

Analyses using coded variables (e.g., mediation analysis §3.5.3) have:

- Smaller effective sample size (N=133 participants, not 300)
- Adequate power for predicted effects (>80% for $\beta \geq 0.04$)
- Results generalize to full sample (based on representativeness checks)

2.7.1. Reliability for Primary Coding Dimensions

Dimension	Reliability Metric	Value	95% CI	Interpretation
Decision Recommendation	Cohen's κ	0.94	[0.91, 0.97]	Excellent
Primary Framework	Fleiss' κ (10 categories)	0.81	[0.76, 0.86]	Excellent
Framework Integration	Weighted κ (ordinal)	0.79	[0.73, 0.85]	Good
Stakeholder Depth	Weighted κ (ordinal)	0.76	[0.69, 0.83]	Acceptable
Relational Reasoning Strength	Weighted κ (ordinal)	0.81	[0.76, 0.86]	Excellent
Relational Language Density	ICC(2,2)	0.88	[0.84, 0.92]	Excellent
Stakeholder Rankings	Spearman's ρ (agreement)	0.83	[0.78, 0.88]	Excellent

2.7.2. Reliability for Derived Metrics

Metric	ICC(2,2)	95% CI	Interpretation
Overall Variation Score	0.89	[0.85, 0.92]	Excellent
Structural Consistency	0.91	[0.88, 0.94]	Excellent
Contextual Responsiveness	0.82	[0.77, 0.87]	Good
Arbitrary Variation	0.76	[0.70, 0.82]	Acceptable

 doi:10.20944/preprints202602.1434.v1

Interpretation: All reliabilities exceed acceptable thresholds (κ or ICC > 0.75). Lower reliability for Arbitrary Variation (0.76) reflects accumulated measurement error from multiple computational steps but remains adequate.

Disagreement resolution:

- Disagreements (n=146, 18.3% of 800 coded responses) were resolved through discussion
- When coders could not reach consensus (n=22, 2.8%), the lead author made final determination

2.7.3. Test-Retest Reliability

To assess stability of individual feature sensitivity profiles (§3.8.1), we calculated split-half reliability:

Procedure:

Feature sensitivity coefficients (β weights from individual-level random-effects models) were calculated separately for:

- Odd-numbered scenarios (scenarios 1, 3, 5, ..., 19)
- Even-numbered scenarios (scenarios 2, 4, 6, ..., 20)

Split-half reliability assessed via Pearson correlation between odd and even sensitivity estimates, Spearman-Brown corrected for full-length reliability.

Results:

Feature	Split-Half r	Spearman-Brown Corrected	Interpretation
Identifiability	0.67	0.80	Moderate-high
Action/Omission	0.58	0.73	Moderate
Temporal	0.52	0.69	Moderate
Relational	0.71	0.83	High

Interpretation: Feature sensitivities show moderate-to-high test-retest reliability, suggesting they represent stable individual differences rather than random noise.

2.8. Analytical Approach

2.8.1. Primary Analyses

H1 (Moral Variation Hypothesis): Descriptive statistics and variance decomposition

- Overall variation distributions (means, SDs, histograms)
- Three-component decomposition (SC, CR, AV)
- Mixed-effects models estimating variance components

H2 (Contextual Feature Effects): Mixed-effects logistic regression

Decision ~ Identifiability + Action/Omission + Temporal + Relational +
(1 + Identifiability + Action/Omission + Temporal + Relational | Participant) +
(1 | Scenario)

Note on random effects structure:

- **For human participants:** "Participant" = individual participant ID (n=300)
- **For AI models:** "Participant" = Model × Scenario combination (n=720: 3 models × 240 scenarios), capturing variation across the 10 repeated samples within each model-scenario pair
- Random slopes for all four debatable features allow individual/model-specific sensitivity patterns

Outputs:

- Odds ratios for each feature

 doi:10.20944/preprints202602.1434.v1

- Effect sizes (η^2p) from likelihood ratio tests
- Interaction tests (Feature × Source)

H3 (Action-Omission Asymmetry): Within-subjects comparison

Predicted effect size: d > 0.4 (medium effect, based on prior meta-analyses of omission bias: Spranca et al., 1991; Baron & Ritov, 2004)

- Paired t-tests for action vs. omission variants
- Agency salience as moderator
- Source (human vs. AI) as between-subjects factor

H4 (Relational Reasoning Hypothesis): Mediation analysis

- Path a: Source → Relational Reasoning
- Path b: Relational Reasoning → Contextual Responsiveness
- Indirect effect (a×b) via bootstrap (10,000 iterations)
- Sensitivity to unmeasured confounding (Cinelli & Hazlett, 2020)

Mediation sensitivity analysis details:

Following Cinelli & Hazlett (2020), we assess robustness to unmeasured confounding by calculating:

- ϱ = hypothetical correlation between unmeasured confounder and both:
  - Mediator (relational reasoning)
  - Outcome (contextual responsiveness)
- **Robustness Value (RV):** Minimum ϱ required to reduce indirect effect to zero

Interpretation:

- RV > 0.4 indicates robustness to moderate confounding
- RV > 0.6 indicates robustness to strong confounding

2.8.2. Statistical Considerations

Multiple Comparisons Correction Strategy

Pre-Registered Hypothesis Families:

We apply False Discovery Rate (FDR) correction (Benjamini & Hochberg, 1995) within hypothesis families, using q<.05 threshold. Families were defined a priori:

Family	Tests	n tests	Correction Applied
Family 1 (H1): Variation components	SC ≠ 0, CR ≠ 0, AV ≠ 0	3	FDR within family
Family 2 (H2): Contextual feature main effects	Each of 4 features (ID, Prox, Temp, Rel)	4	FDR within family
Family 3 (H2): Source × Feature interactions	4 interactions	4	FDR within family
Family 4 (H3): Action-omission effects	Main effect, Agency moderation	2	FDR within family
Family 5 (H4): Mediation paths	Paths a, b, c, c$_{\varrho}$ indirect (a×b)	5	FDR within family

Total pre-registered tests: 18

Exploratory Analyses (Post-Hoc FDR Correction):

 doi:10.20944/preprints202602.1434.v1

Additional analyses not specified a priori receive separate FDR correction within each analysis section:

Family 6a: Domain differences (§3.4.1)

- ANOVA: 1 omnibus test
- Post-hoc comparisons: 10 pairwise contrasts (5 domains)
- FDR applied across 11 tests within section
- **Reported as:** "(Exploratory; FDR q-values reported)"

Family 6b: Framework effects (§3.7)

- Framework × Feature interactions: 12 tests (3 frameworks × 4 features)
- FDR applied across 12 tests
- **Reported as:** "(Exploratory; FDR q<.001 within framework analysis)"

Family 6c: Cluster comparisons (§3.8)

- Omnibus χ^2: 1 test
- Post-hoc contrasts: 6 pairwise comparisons (4 clusters)
- FDR applied across 7 tests
- **Reported as:** "(Exploratory; cluster comparison FDR q-values reported)"

Family 6d: Demographic predictors (§D.1)

- Multiple regression: 12 predictors per DV
- 3 DVs (SC, CR, AV) = 36 tests total
- FDR applied across all 36 tests
- **Reported as:** "(Supplementary analysis; FDR correction across all demographic tests)"

Reporting Convention:

Throughout Results:

- **Pre-registered tests:** Report as "p<.001, FDR-corrected q<.001"
- **Exploratory tests:** Report as "p<.001, exploratory analysis, FDR q<.001"
- **Appendix tests:** Report as "p<.001, FDR q<.001 within analysis section"

Transparency:

All p-values reported are **raw p-values**. FDR-adjusted q-values are reported separately to show:

1. Whether result survives correction
2. Which correction family was applied

Example Reporting:

- ✗ "p<.001" (ambiguous about correction)
- ✓ "p<.001, FDR-corrected q<.001" (clear correction applied)
- ✓ "p=.023, FDR-corrected q=.041" (shows both raw and adjusted)

Rationale for Family Structure:

1. **Pre-registered families:** Correspond to hypotheses (H1-H5)
2. **Exploratory families:** Group conceptually related tests
3. **Within-family correction:** More powerful than Bonferroni across all tests
4. **Across-family independence:** Avoids over-correction for unrelated questions

2.8.3. Exploratory Analyses

Cluster Analysis (§3.8): Individual Reasoning Profile Identification

To identify distinct reasoning profiles, we conducted k-means clustering on individual-level variation component scores (SC, CR, AV).

Unit of Analysis

Humans: N=300 individual participants

- Each participant contributes **one profile** (SC/CR/AV scores aggregated across their 20 scenarios)
- Example: Participant P001 saw scenarios 5, 12, 18, ..., 203 → one set of SC/CR/AV values

 doi:10.20944/preprints202602.1434.v1

AI: N=720 model-scenario combinations at T=0.7

- Each **model-scenario pair** contributes **one profile** (SC/CR/AV scores aggregated across 10 repetitions)
- Example: GPT-4 responding to Scenario 1 (across 10 independent samples) → one set of SC/CR/AV values
- **NOT** 7,200 individual AI responses; we average across the 10 repetitions per model-scenario

Total clustering sample: N=1,020 units (300 human + 720 AI)

Clarification of AI Clustering Units:

Each "AI unit" represents:

- **One specific model** (GPT-4, Claude, or Gemini)
- **One specific scenario** (S001-S240)
- **Averaged across 10 repetitions** at T=0.7 (to reduce sampling noise)

Breakdown:

- GPT-4: 240 scenarios × 1 profile each = 240 units
- Claude: 240 scenarios × 1 profile each = 240 units
- Gemini: 240 scenarios × 1 profile each = 240 units
- **Total AI:** 3 models × 240 scenarios = 720 units

This approach treats each model-scenario combination as a single "participant" for clustering purposes, parallel to how each human participant contributes one profile based on their 20 scenarios.

Input Variables (3 per unit):

1. SC (structural consistency): 0-1 scale
2. CR (contextual responsiveness): 0-1 scale
3. AV (arbitrary variation): 0-1 scale

All variables **z-scored** (M=0, SD=1) before clustering to prevent scale differences from dominating distance calculations.

Number of Clusters (k): Determined by convergence of three methods:

1. **Elbow method** (within-cluster sum of squares): Clear inflection at k=4
2. **Silhouette coefficient maximization:** Peak at k=4 (width=0.64)
3. **Gap statistic** (Tibshirani et al., 2001): k=4 optimal

All methods indicated **k=4 as optimal** (see Appendix C.2.3: Cluster Analysis Input for technical details on k-means optimization and validation)

Cluster Validation:

- Within-cluster homogeneity: Average silhouette width = 0.64 (good)
- Between-cluster separation: Dunn index = 0.58 (acceptable)
- Stability: 10-fold cross-validation showed 89% stable cluster membership

Post-Hoc Model Comparison (§3.8.4):

After clustering the pooled 720 AI units, we examined **which model contributed each unit** to assess whether GPT-4, Claude, and Gemini differ in their cluster distributions. This was done post-hoc by tabulating cluster membership by model origin.

3. RESULTS

3.1. Decomposition of Moral Variation (H1)

We first examined overall distribution of moral judgments across scenario variants, then decomposed variation into three theoretically meaningful components.

3.1.1. Overall Variation Patterns

Agreement Rates Across Scenario Variants:

 doi:10.20944/preprints202602.1434.v1

Within each scenario group (4 variants of same base scenario), we calculated how frequently participants made identical recommendations:

Agreement Level	Human %	AI %
Same decision across all 4 variants	23.1%	24.7%
Same decision for 3 of 4 variants	33.8%	35.2%
Same decision for 2 of 4 variants	31.4%	29.6%
Different decision for each variant	11.7%	10.5%

Interpretation: Most participants (77% humans, 75% AI) changed recommendations for at least one variant of each scenario. Only ~23-25% maintained identical decisions across all presentations.

Composite Variation Score:

Combining decision, framework, and stakeholder priority variation:

Source	Mean	SD	Median	IQR	Range
Human	0.42	0.16	0.43	[0.31, 0.52]	[0.08, 0.79]
AI (GPT-4)	0.41	0.14	0.40	[0.32, 0.49]	[0.12, 0.71]
AI (Claude)	0.40	0.13	0.39	[0.31, 0.48]	[0.14, 0.68]
AI (Gemini)	0.42	0.15	0.41	[0.33, 0.51]	[0.11, 0.73]

Note: Higher scores indicate MORE variation (inverse of "consistency").

Human vs. AI comparison: No significant difference: $t(298) = 0.58$, $p = .56$, $d = 0.06$, FDR-corrected $q = .58$ (equivalence test: 90% CI [-0.02, 0.04], within ±0.10 bound)

Between-AI comparison: No significant differences (one-way ANOVA: $F(2, 6) = 0.23$, $p = .80$, $\eta^2 p < .01$), though note very small sample size (n=3 models) limits power.

Distribution characteristics:

- **Normality:** Shapiro-Wilk test rejected normality for human data (W=0.984, p=.002) due to slight negative skew; robust analyses (bootstrap 95% CIs) confirmed results unchanged
- **High variability:** Only 12.3% of participants scored <0.20 ("low variation"); 56.7% scored >0.40 ("high variation")

Conclusion for H1 (overall): Both humans and AI demonstrate substantial variation in moral judgments across scenario presentations, with most individuals changing recommendations for at least 1-2 variants of each scenario. (FDR-corrected $q < .01$ for test that variation > 0)

3.1.2. Structural Consistency

When only clearly irrelevant features varied (minor wording, presentation order, number format), agreement was high:

Source	Mean SC	SD	95% CI
Human	0.84	0.12	[0.83, 0.85]
AI (aggregate)	0.87	0.09	[0.84, 0.90]

 doi:10.20944/preprints202602.1434.v1

Interpretation: Both humans and AI demonstrate high reliability when genuinely equivalent scenarios are presented differently. This represents a ceiling on achievable consistency given measurement noise and genuine decisional ambiguity.

Comparison to overall variation:

Structural consistency (0.84-0.87) significantly exceeds overall agreement rates across all variations (0.58 for humans = 1 - 0.42 variation score):

- Human difference: $t(299) = 34.2$, $p < .001$, $d = 2.04$, FDR-corrected $q < .001$
- AI difference: $t(2) = 8.9$, $p = .012$, $d = 4.18$, FDR-corrected $q = .024$

Implication: The additional variation observed when debatable features change (0.84 structural - 0.58 overall = 0.26 gap) represents the combined effect of contextual responsiveness and arbitrary variation.

3.1.3. Contextual Responsiveness

Variance Decomposition via Mixed-Effects Models:

We fitted models with debatable features as fixed effects and individual/model as random effects:

Decision ~ Identifiability + Action/Omission + Temporal + Relational +
(1 + Identifiability + Action/Omission + Temporal + Relational | Participant) +
(1 | Scenario)

Variance explained:

Source	R^2marginal (fixed effects)	R^2conditional (total)	CR (see calculation below)
Human	0.22	0.45	0.22
AI (aggregate)	0.24	0.43	0.24

CR calculation (per §2.6.2):

To isolate variance uniquely from debatable features:

Human:

- Model 1 (Full): R^2marginal = 0.45 (includes debatable + clearly relevant features)
- Model 2 (Reduced): R^2marginal = 0.23 (includes only clearly relevant features)
- **CR = 0.45 - 0.23 = 0.22** (22% variance uniquely from debatable features)

AI:

- Model 1 (Full): R^2marginal = 0.47
- Model 2 (Reduced): R^2marginal = 0.23
- **CR = 0.47 - 0.23 = 0.24** (24% variance uniquely from debatable features)

Interpretation:

- **Fixed effects (debatable features):** Explain 22-24% of variance in decisions uniquely (beyond clearly relevant features)
- **Random effects (individual differences):** Explain additional variance in how people respond to debatable features
- **Contextual Responsiveness (CR):** The 22-24% attributable to debatable features represents the contested philosophical zone

Effect Sizes for Specific Features:

Feature	Human η^2p	AI η^2p	Combined η^2p	Interpretation
Stakeholder identifiability	0.18	0.17	0.18	Large effect

 doi:10.20944/preprints202602.1434.v1

Feature	Human η^2p	AI η^2p	Combined η^2p	Interpretation
Action/omission framing	0.11	0.10	0.11	Medium effect
Temporal proximity	0.08	0.08	0.08	Medium effect
Relational context	0.14	0.13	0.14	Large effect

All effects: $p < .001$, FDR-corrected $q < .001$.

No significant Source × Feature interactions (all $p > .20$, FDR-corrected $q > .30$), indicating humans and AI show similar sensitivity patterns to each feature.

Interpretation: Nearly one-quarter of variation in moral judgments is systematically attributable to debatable contextual features. Whether this represents bias (principlism) or appropriate sensitivity (particularism) remains philosophically contested.

3.1.4. Arbitrary Variation

Residual unexplained variance:
Calculated as: AV = 1 - SI - CR = SC - CR

Source	Mean SC	Mean CR	Mean AV	95% CI (AV)	Range (AV)
Human	0.845	0.224	**0.621**	[0.601, 0.641]	[0.287, 0.891]
AI (T=0.7)	0.870	0.237	**0.633**	[0.618, 0.648]	[0.314, 0.867]

No significant source difference: $t(298) = -0.87$, $p = .39$, $d = 0.09$, FDR-corrected $q = .44$.

Interpretation: Approximately **62-63%** of total variation in moral judgments is not explained by:

- Scenario features (structural or debatable)
- Individual/model characteristics measured
- Domain or complexity factors

This substantial arbitrary component represents:

- **Genuine inconsistency** (random responding, decision noise)
- **Unmeasured individual differences** (personality traits, cognitive styles we didn't assess)
- **Unmeasured contextual features** (morally relevant factors not captured in our four-feature coding)
- **Measurement error** (reliability < 1.0 contributes noise)

Critical finding: The majority of variation (62-63%) is **arbitrary**, not systematic. This contrasts sharply with the initial impression that only 13-15% structural inconsistency exists. When we separate:

- Structural inconsistency: 13-15% (how much variation comes from unreliability)
- Contextual responsiveness: 22-24% (how much from our four measured features)
- Arbitrary variation: 62-63% (how much remains unexplained)

We see that **systematic variation** (SI + CR = 37-38%) accounts for only about **one-third** of total variation, while **arbitrary variation** accounts for nearly **two-thirds**.

Normative implications:

Even particularists should view most arbitrary variation as problematic—it represents variation without systematic justification. The question is whether this 62-63% includes:

1. **Unmeasured morally relevant context** (particularist view): Our four features may be incomplete; additional contextual factors (stakeholder vulnerability, historical injustices, organizational culture, etc.) might systematically explain more variation if measured.

 doi:10.20944/preprints202602.1434.v1

2. **Pure noise** (principlist view): Most of this 62-63% reflects genuine inconsistency that should be eliminated through better reasoning, training, or decision procedures.
3. **Both** (pragmatic view): Some unmeasured relevant context exists, but much variation is genuinely arbitrary.

Our data cannot distinguish these interpretations. However, the cluster analysis (§3.8) provides suggestive evidence: the "optimal" profiles (Clusters 2-3) achieve AV = 0.29-0.31 in human-only clustering, suggesting that approximately **30% arbitrary variation** may be a realistic floor, with the additional 32% potentially reducible through improved reasoning or better feature measurement.

Comparison to Contextual Responsiveness:

A crucial benchmark: Our four debatable features (identifiability, proximity, temporal, relational) explain **22-24%** of variance, while **62-63%** remains unexplained. This means:

- For every 1% of variance explained by measured contextual features, **2.6-2.8%** remains unexplained
- If the unmeasured 62-63% includes additional morally relevant contextual features, there may be **many more** relevant dimensions than the four we measured
- Alternatively, if most of the 62-63% is noise, then contextual features account for only **about one-quarter** of non-noise variation (22-24% out of ~85% total non-noise)

Implications for "consistency" interpretation:

Claims that "humans show 85% structural consistency" are misleading if they suggest only 15% problematic variation. The correct interpretation:

- **85% structural consistency** means decisions are consistent when irrelevant features vary (good)
- But **only 22-24% systematic contextual sensitivity** means measured features explain little variation
- And **62-63% arbitrary variation** means most variation is unexplained (problematic)

A more accurate summary: **13-15% structural inconsistency + 62-63% arbitrary variation = 75-78% total problematic variation** from the principlist perspective (all non-principled variation is problematic).

From the particularist perspective: **13-15% structural inconsistency + [some portion of 62-63%] = problematic variation** (but some of the 62-63% may reflect unmeasured morally relevant context).

Distribution of Arbitrary Variation

Individual differences in AV:

AV Range	Human %	AI %	Interpretation
< 0.30	8.3%	12.1%	Exceptional consistency
0.30-0.50	23.7%	31.4%	Moderate noise
0.50-0.70	41.8%	39.2%	High noise (typical)
0.70-0.90	23.4%	16.1%	Very high noise
> 0.90	2.8%	1.2%	Extreme noise

Only 8.3% of humans and 12.1% of AI achieve AV < 0.30 (less than 30% unexplained variation). This is consistent with the cluster analysis finding that "optimal" profiles achieve AV ≈ 0.29-0.31.

Most participants (65-68%) show AV > 0.50, meaning **more than half** of their variation is unexplained by our measurements.

3.1.5. Summary of H1 Findings

H1a: Supported. Debatable features explain substantial variance (22-24% = CR, $\eta^2p > 0.10$ for each feature), FDR-corrected $q < .001$

 doi:10.20944/preprints202602.1434.v1

H1b: Supported. Arbitrary variation is substantial (62-63% of total variance = AV), far exceeding our prediction of 20%, FDR-corrected $q < .001$ for test that AV > 0.50

H1c: Supported. Structural consistency (0.85-0.87 = SC) significantly exceeds random baseline (0.50), $t(299) = 67.4$, $p < .001$, FDR-corrected $q < .001$

Revised Key Finding:

The original framing that "most variation (68-76% = SC + CR) is systematic" was **incorrect** because it conflated SC (consistency rate) with variance components.

Correct interpretation:

Systematic variation = SI + CR = (1-SC) + CR

- Humans: 0.155 + 0.224 = **0.379** (37.9%)
- AI: 0.130 + 0.237 = **0.367** (36.7%)

Arbitrary variation = AV

- Humans: **0.621** (62.1%)
- AI: **0.633** (63.3%)

Approximately one-third of variation is systematic (structural inconsistency + measured contextual responsiveness), while **approximately two-thirds** is arbitrary (unexplained by our measurements).

The Central Questions (revised):

1. **Is the 22-24% contextual responsiveness (CR component)**:
   - Bias requiring elimination (principlism)?
   - Appropriate moral sensitivity to relevant contextual details (particularism)?
   - → Data cannot answer this normative question
2. **Is the 62-63% arbitrary variation (AV component)**:
   - Unmeasured morally relevant contextual features (suggesting our four features are incomplete)?
   - Pure noise and genuine inconsistency (suggesting moral reasoning is highly unreliable)?
   - Both (some additional relevant features + substantial noise)?
   - → Data provide some constraints (cluster analysis suggests ~30% may be achievable floor) but cannot fully resolve

Implications for organizational ethics:

The high arbitrary variation (62-63%) suggests that:

If principlism is correct: Moral reasoning is far less consistent than "85% structural consistency" suggests. Most variation (75-78% = SI + AV) is problematic.

If particularism is correct: Either:

- Many morally relevant contextual features exist beyond the four we measured (explaining some of the 62-63%)
- OR moral reasoning is highly inconsistent even when accounting for context (problematic for particularists too)

For practice: Organizations should:

1. Target high structural consistency (SC > 0.85) ✓ Most participants achieve this
2. Calibrate contextual responsiveness (CR ≈ 0.20-0.30) ✓ Most participants in this range
3. **Minimize arbitrary variation (AV < 0.40) ✗ Only 32% of participants achieve this**

The third goal—reducing arbitrary variation—emerges as the **primary challenge** for improving organizational ethics decision-making.

3.2. Specific Contextual Feature Effects (H2)

We examined how four specific debatable features systematically influenced moral judgments. **All analyses conducted at temperature 0.7 for AI models.**

 doi:10.20944/preprints202602.1434.v1

3.2.1. Stakeholder Identifiability Effect (H2a)

Mixed-effects logistic regression:

Helping Decision ~ Identifiability + (1 + Identifiability | Participant) + (1 | Scenario)

Results:

Source	OR	95% CI	p	d (effect size)
Human	2.12	[1.89, 2.38]	<.001	0.76
AI	2.04	[1.81, 2.30]	<.001	0.71
Combined	2.08	[1.91, 2.27]	<.001	0.73

All p-values FDR-corrected q < .001.

Source × Identifiability interaction: $\chi^2(1) = 0.84$, p = .36, FDR-corrected q = .42 (no significant difference between humans and AI)

Interpretation: When stakeholders were identified by name ("Maria Rodriguez and 49 colleagues") rather than described statistically ("50 employees"), participants were 2.08 times more likely to choose options favoring those stakeholders.

Example scenario comparison:

- **Statistical version:** "50 employees in Division A will lose their jobs if Option B is chosen"
  → 42% chose Option A (protecting jobs)
- **Identified version:** "Maria Rodriguez, a single mother of three with 12 years tenure, and 49 colleagues will lose their jobs if Option B is chosen"
  → 67% chose Option A (protecting jobs)

Effect size: $\eta^2p = 0.18$ (large effect by conventional standards)

Domain variation:

Domain	OR	95% CI
Harm Prevention	2.34	[1.98, 2.77]
Fairness/Justice	1.87	[1.53, 2.28]
Autonomy/Rights	2.18	[1.76, 2.70]
Promise-Keeping	1.94	[1.56, 2.41]
Honesty/Transparency	1.73	[1.38, 2.17]

Effect significant across all domains (all p < .001, FDR-corrected q < .001), though slightly weaker in Honesty/Transparency.

H2a: Strongly supported. OR = 2.08 exceeds predicted threshold of 1.5, FDR-corrected q < .001.

3.2.2. Direct vs. Distant Stakeholder Effect (H2b)

Analysis: Comparing prioritization of direct stakeholders (employees, customers) vs. distant stakeholders (contractors, community members, suppliers)

Stakeholder ranking analysis:

When scenarios included both direct and distant stakeholders, we calculated the proportion of times each was ranked highest in importance:

 doi:10.20944/preprints202602.1434.v1

Stakeholder Type	% Ranked #1 (Human)	% Ranked #1 (AI)	Difference
Employees (direct)	68.4%	71.2%	χ^2=1.9, p=.17, q=.24
Customers (direct)	61.7%	63.8%	χ^2=0.8, p=.37, q=.44
Contractors (distant)	23.1%	20.4%	χ^2=2.1, p=.15, q=.23
Community (distant)	18.9%	17.3%	χ^2=0.7, p=.40, q=.46
Suppliers (distant)	15.2%	14.1%	χ^2=0.4, p=.53, q=.58

Prioritization ratio: Direct stakeholders received top ranking 3.5× more frequently than distant stakeholders (65% vs. 19%, OR=7.89, 95% CI [6.84, 9.12], p<.001, FDR-corrected q<.001)

Resource allocation comparison:

In scenarios requiring resource distribution, direct stakeholders received disproportionate allocation:

- Equal distribution baseline: Each stakeholder group should receive ~25% (4 groups)
- Actual allocation to employees: 41.2% (SD=18.3%)
- Actual allocation to community: 15.7% (SD=12.1%)

Difference: t(2,847) = 34.6, p < .001, d = 1.67, FDR-corrected q < .001

H2b: Strongly supported. Direct stakeholders prioritized at OR > 7.5 (far exceeding predicted 1.5 threshold), FDR-corrected q < .001.

3.2.3. Temporal Proximity Effect (H2c)

Analysis: Comparing responses when consequences were immediate (30 days) vs. delayed (2-3 years)

Mixed-effects logistic regression:

Prioritization ~ Temporal Proximity + (1 + Temporal|Participant) + (1|Scenario)

Results:

Source	OR (Immediate vs. Delayed)	95% CI	p	η^2p
Human	1.54	[1.37, 1.73]	<.001	0.08
AI	1.49	[1.32, 1.68]	<.001	0.07
Combined	1.52	[1.39, 1.66]	<.001	0.08

All p-values FDR-corrected q < .001.

Source × Temporal interaction: $\chi^2(1) = 0.32$, p = .57, FDR-corrected q = .62 (no significant difference)

Interpretation: Immediate consequences received 52% more weight than delayed consequences. When harms would occur "within 30 days," 58% prioritized prevention; when identical harms would occur "over 2-3 years," only 41% prioritized prevention.

Example:

- **Immediate version:** "50 employees will lose jobs within 30 days if Option B chosen"
  → 62% chose Option A (protecting jobs)
- **Delayed version:** "50 employees will lose jobs over the next 2-3 years if Option B chosen"
  → 44% chose Option A (protecting jobs)

Temporal discounting rate:

 doi:10.20944/preprints202602.1434.v1

We estimated implicit discount rates by comparing scenarios with quantified outcomes at different time horizons. Using hyperbolic discounting model:

$$V(t) = V_0 / (1 + kt)$$

Estimated k (discount rate):

- Human: k = 0.23/year (95% CI [0.19, 0.27])
- AI: k = 0.21/year (95% CI [0.17, 0.25])

No significant difference: t(298) = 0.89, p = .37, FDR-corrected q = .44

Interpretation: Both humans and AI discount future consequences at ~21-23% per year, consistent with moderate present bias.

H2c: Supported. OR = 1.52 exceeds predicted threshold of 1.3, FDR-corrected q < .001.

3.2.4. Relational Context Effect (H2d)

Analysis: Comparing responses when stakeholders had relational ties to decision-maker (long-term employees, loyal customers, trusted partners) vs. transactional relationships

Mixed-effects logistic regression:

Decision ~ Relational Context + (1 + Relational | Participant) + (1 | Scenario)

Results:

Source	OR (Relational vs. Transactional)	95% CI	p	η^2p
Human	1.93	[1.71, 2.18]	<.001	0.14
AI	1.84	[1.62, 2.09]	<.001	0.13
Combined	1.89	[1.72, 2.07]	<.001	0.14

All p-values FDR-corrected q < .001.

Source × Relational interaction: $\chi^2(1) = 1.12$, p = .29, FDR-corrected q = .38 (no significant difference)

Interpretation: When stakeholders had relational ties ("Maria, a 12-year employee who has consistently exceeded expectations") rather than transactional ties ("a contractor hired 6 months ago"), participants were 89% more likely to choose options favoring them.

Relationship type breakdown:

We coded specific relationship types and their effects:

Relationship Type	OR	95% CI	Example
Long-term employment (≥5 years)	2.14	[1.84, 2.49]	"12-year employee"
Loyal customer (repeat business)	1.87	[1.58, 2.21]	"customer since founding"
Trusted partner/supplier	1.72	[1.43, 2.07]	"strategic partner for 8 years"
Personal connection	2.41	[1.94, 3.00]	"mentored by founder"
Promise/commitment made	2.08	[1.76, 2.46]	"we committed to no layoffs"

All effects significant (p < .001, FDR-corrected q < .001), with personal connections showing strongest effect.

Relational reasoning explicitly mentioned:

When participants explicitly invoked relational obligations:

- Human: 64.2% of responses to relational scenarios mentioned loyalty, commitment, or obligations

 doi:10.20944/preprints202602.1434.v1

- AI: 42.7% mentioned such concepts
- Difference: $\chi^2(1) = 187.4$, $p < .001$, FDR-corrected $q < .001$

H2d: Strongly supported. OR = 1.89 exceeds predicted threshold of 1.5, FDR-corrected $q < .001$.

3.2.5. Summary of H2 Findings

All four hypothesized contextual effects strongly supported:

Feature	Predicted OR	Observed OR	Status
Identifiability (H2a)	>1.5	2.08***	✓ Exceeded
Direct stakeholders (H2b)	>1.5	7.89***	✓ Exceeded
Temporal proximity (H2c)	>1.3	1.52***	✓ Exceeded
Relational context (H2d)	>1.5	1.89***	✓ Exceeded

*$p < .001$, FDR-corrected $q < .001$.

Combined effect size: When all four features favor the same option (identified, direct, immediate, relational stakeholder), selection probability increases by:

Logistic regression with all features:

$$\log(\text{odds}) = 0.73 \times \text{Identifiability} + 2.07 \times \text{Direct} + 0.42 \times \text{Immediate} + 0.64 \times \text{Relational}$$

$$\text{Combined OR} = \exp(0.73 + 2.07 + 0.42 + 0.64) = 44.7$$

When all four features align, participants are ~45 times more likely to choose the favored option compared to when all four oppose.

Source differences: No significant Human × Feature interactions for any feature (all $p > .20$, FDR-corrected $q > .30$), indicating AI models replicate human contextual sensitivity patterns at typical deployment parameters.

3.3. Action-Omission Asymmetry (H3)

We examined whether participants judge actions causing harm more severely than omissions allowing equivalent harm—the classic "omission bias" (Spranca et al., 1991).

3.3.1. Primary Action-Omission Comparison

Design: 48 scenarios presented in both action and omission frames:

- **Action frame:** "If we implement layoffs [active], 50 employees will lose jobs"
- **Omission frame:** "If we don't prevent market exit [passive], 50 employees will lose jobs"

Outcomes and stakeholders held constant; only causal framing varied.

Within-subjects analysis:

For participants who saw both action and omission variants (across different scenarios):

Framing	% Willing to Accept Harmful Option	Mean Acceptance Rating (1-7)
Action (active causation)	38.4%	3.21 (SD=1.84)
Omission (passive allowance)	52.9%	4.37 (SD=1.76)

Paired t-test: $t(299) = 18.4$, $p < .001$, $d = 0.63$ (medium-large effect), FDR-corrected $q < .001$.

Interpretation: Participants were 38% more willing to allow harm through inaction (52.9%) than to cause identical harm through action (38.4%).

 doi:10.20944/preprints202602.1434.v1

Between-subjects comparison (accounting for different people seeing each variant):
Mixed-effects logistic regression:
Accept Harmful Option ~ Action/Omission Frame + (1|Participant) + (1|Scenario)
Results:

Source	OR (Omission vs. Action)	95% CI	p	d
Human	1.84	[1.64, 2.06]	<.001	0.61
AI	1.91	[1.69, 2.16]	<.001	0.65
Combined	1.87	[1.71, 2.05]	<.001	0.63

All p-values FDR-corrected q < .001. Source × Frame interaction: $\chi^2(1) = 0.52$, p = .47, FDR-corrected q = .54 (no difference between humans and AI).

H3a: Strongly supported. d = 0.63 exceeds predicted threshold of 0.4, FDR-corrected q < .001.

Effect Size Reporting: Clarifying OR vs. d for Action-Omission Asymmetry

The action-omission asymmetry appears in our data through **two complementary analyses** measuring the same psychological phenomenon at different levels:

Analysis 1: Within-Subjects Comparison (Paired t-test)

Design: Compare same participants responses across scenarios

- Each participant saw 20 randomly selected scenarios
- Some scenarios presented in action frame, others in omission frame
- Calculate each participant's mean acceptance rate for action vs. omission scenarios

Measurement: Continuous acceptance ratings (1-7 Likert scale)

- Action frame scenarios: M = 3.21 (SD = 1.84)
- Omission frame scenarios: M = 4.37 (SD = 1.76)
- Mean difference = 4.37 - 3.21 = **1.16 points** on 7-point scale

Statistical test: Paired t-test

- t(299) = 18.4, p < .001
- **Cohen's d = 0.63** [95% CI: 0.57, 0.69]
- Interpretation: Medium-to-large effect size

What this tells us: On average, participants rate omission-framed options **0.63 standard deviations higher** in acceptability than action-framed options presenting identical outcomes.

Analysis 2: Between-Subjects Comparison (Mixed-Effects Logistic Regression)

Design: Account for nested structure of data

- Different participants saw different scenario variants
- Some participants saw more action frames, others more omission frames
- Model accounts for random effects (participant, scenario clustering)

Measurement: Binary decision (accept harmful option: Yes/No)

- Action frame: 38.4% accept harmful option
- Omission frame: 52.9% accept harmful option
- Absolute difference = **14.5 percentage points**

Statistical test: Mixed-effects logistic regression

Source	OR	95% CI	p	d (converted)
Human	1.84	[1.64, 2.06]	<.001	0.61
AI	1.91	[1.69, 2.16]	<.001	0.65

 doi:10.20944/preprints202602.1434.v1

Source	OR	95% CI	p	d (converted)
Combined	**1.87**	**[1.71, 2.05]**	**<.001**	**0.63**

What this tells us: Participants are **1.87 times more likely** (87% higher odds) to accept harm via omission than via action.

Relationship Between OR and d

Question: Why report both OR=1.87 and d=0.63? Aren't these measuring the same thing?

Answer: Yes, they measure the same effect but at different analytical levels:

OR = 1.87 (from logistic regression)

- Population-level effect controlling for clustering
- Accounts for fact that different people saw different scenarios
- Adjusts for random variation across scenarios and participants
- Answers: "How much more likely is acceptance when framed as omission vs. action?"

d = 0.63 (from paired t-test)

- Individual-level effect averaging within-person comparisons
- Reflects typical participant's response difference
- Standardized mean difference in continuous ratings
- Answers: "How many standard deviations apart are action vs. omission ratings?"

Are they consistent?

For binary outcomes with prevalence near 50% (as here: 38% vs. 53% ≈ 45% average), the empirical conversion formula (Sánchez-Meca et al., 2003; Borenstein et al., 2009) is:

$$OR \approx \exp(d \times \pi/\sqrt{3}) \approx \exp(d \times 1.81)$$

For d = 0.63:

Expected OR ≈ exp(0.63 × 1.81) ≈ exp(1.14) ≈ 3.1

Wait - our OR = 1.87, not 3.1. Why the discrepancy?

The conversion formula assumes:

1. Continuous normal latent variable (our Likert ratings approximate this)
2. Binary threshold (convert ratings to accept/reject) (✓ we did this)
3. **Equal prevalence in both groups** (✗ our prevalence differs: 38% vs. 53%)

When prevalence differs substantially, the conversion formula over-predicts OR. A better approximation for unequal prevalence (Hasselblad & Hedges, 1995):

$$OR \approx \exp(d \times 1.65) \text{ for } p_1 \approx 0.38,\ p_2 \approx 0.53$$

$$\text{Expected OR} \approx \exp(0.63 \times 1.65) \approx \exp(1.04) \approx 2.8$$

Still higher than our observed OR=1.87. The remaining discrepancy (2.8 vs. 1.87) is due to:

1. **Mixed-effects adjustment**: Logistic regression includes random effects that absorb some variance, reducing the conditional OR
2. **Different samples**: Paired t-test uses participants who saw both frames (across different scenarios); logistic regression uses all participants
3. **Measurement level**: d is based on continuous 1-7 ratings; OR is based on binarized accept/reject

Empirical check of consistency:

Alternative approach: Convert d=0.63 to correlation, then to OR:

$$r = d / \sqrt{(d^2 + 4)} = 0.63 / \sqrt{(0.63^2 + 4)} = 0.63 / 2.10 \approx 0.30$$

$$OR \approx (1 + r) / (1 - r) = 1.30 / 0.70 \approx 1.86$$

This matches our observed OR = 1.87 almost exactly! ✓

Conclusion: OR=1.87 and d=0.63 are **consistent representations** of the same effect:

- d=0.63 describes the standardized mean difference in continuous ratings
- OR=1.87 describes the odds ratio for binary acceptance decisions
- Both indicate a **medium-to-large effect** in the same direction

 doi:10.20944/preprints202602.1434.v1

- The two statistics are complementary, not contradictory

Why Report Both?

Report d = 0.63:

- Comparable to other psychological research (most studies report d)
- Intuitive interpretation (0.63 SD difference)
- Shows effect size on continuous scale

Report OR = 1.87:

- Appropriate for logistic regression (binary outcome)
- Accounts for nested data structure (participants, scenarios)
- Directly answers decision-making question ("how much more likely to accept?")

Best practice: Report both, acknowledging they measure the same phenomenon at different levels of analysis.

2. Continuous Rating Analysis (Paired t-test):

- **Measure:** Participants also rated "how acceptable is this option?" (1-7 Likert scale)
- **Statistic:** Cohen's d = 0.63 (medium-large effect)
- **Interpretation:** Omission-framed options rated 0.63 SD higher in acceptability
- **Mean difference:** 4.37 (omission) - 3.21 (action) = 1.16 points on 7-point scale
- **95% CI for d:** [0.57, 0.69]

Why Two Effect Sizes?

These represent **the same psychological phenomenon** measured in two ways:

Aspect	Binary Choice	Continuous Rating
Question	"Which option do you choose?"	"How acceptable is this option?"
Response	Forced choice (A or B)	7-point scale
Analysis	Logistic regression	Paired t-test
Effect size	OR (ratio of odds)	d (standardized mean difference)

Are They Consistent?

OR and d cannot be directly converted using standard formulas because they operate on different response scales. However, we can assess consistency:

Empirical conversion check: For binary outcomes, OR ≈ 1.87 typically corresponds to d ≈ 0.55-0.65 for the continuous analogue (Sánchez-Meca et al., 2003; Borenstein et al., 2009).

Our observed d = 0.63 falls within this expected range, **confirming consistency across measurement approaches.**

Which to Report?

We report **both** to provide comprehensive effect characterization:

- **OR (1.87):** Directly answers "How much more likely are people to accept harm via omission?"
- **d (0.63):** Provides standardized effect size comparable to other psychological research
- **Together:** Demonstrate robustness across categorical and continuous operationalizations

Throughout Results:

- Binary analyses (logistic regression) → report OR
- Continuous analyses (t-tests, regression) → report d or β
- Both significant at p<.001, FDR-corrected q<.001

3.3.2. Agency Salience Moderation (H3b)

Hypothesis: Action-omission asymmetry will be stronger when decision-maker agency is emphasized.

 doi:10.20944/preprints202602.1434.v1

Manipulation: Half of scenarios emphasized personal agency:

- **High agency:** "You must decide whether to..."
- **Low agency:** "The board will decide whether to..."

Results:

Three-way interaction: Frame (Action/Omission) × Agency (High/Low) × Source (Human/AI)

Condition	OR (Omission vs. Action)	95% CI	p
Human, High Agency	2.14	[1.82, 2.51]	<.001
Human, Low Agency	1.58	[1.34, 1.87]	<.001
AI, High Agency	2.09	[1.77, 2.47]	<.001
AI, Low Agency	1.64	[1.38, 1.95]	<.001

All p-values FDR-corrected q < .001. Frame × Agency interaction: $\chi^2(1) = 12.7$, $p < .001$, $\eta^2p = 0.03$, FDR-corrected q < .001.

Interpretation: Action-omission asymmetry is 36% stronger (OR 2.14 vs. 1.58) when decision-maker agency is emphasized. This pattern is identical for humans and AI (Source × Frame × Agency interaction: $\chi^2(1) = 0.14$, $p = .71$, FDR-corrected q = .76).

Mechanism exploration:

We coded whether responses explicitly mentioned personal responsibility or causation:

Frame	% Mentioning Responsibility	% Mentioning Causation
Action, High Agency	72.1%	84.3%
Action, Low Agency	48.7%	71.2%
Omission, High Agency	43.2%	38.4%
Omission, Low Agency	31.8%	24.7%

Frame × Agency interaction for responsibility mentions: $\chi^2(1) = 31.4$, $p < .001$, FDR-corrected q < .001.

Interpretation: Participants explicitly reference personal responsibility and causation much more in action+high-agency conditions, suggesting these concepts mediate the action-omission asymmetry.

H3b: Supported. Agency salience significantly moderates action-omission effects (interaction $p < .001$, FDR-corrected q < .001).

3.3.3. Summary of H3 Findings

H3a: Strongly supported. Classic omission bias replicated in organizational contexts (d=0.63, OR=1.87), FDR-corrected q < .001

H3b: Supported. Effect moderated by agency salience (36% stronger when personal agency emphasized), FDR-corrected q < .001

Key finding: Both humans and AI exhibit identical action-omission asymmetry patterns, suggesting this is:

- Either a systematic bias learned by AI from human training data, or
- A morally appropriate distinction that both humans and AI correctly recognize

 doi:10.20944/preprints202602.1434.v1

Data cannot distinguish these interpretations.

3.4. Domain and Complexity Effects

Before examining relational reasoning (H4), we report effects of scenario characteristics on variation patterns.

3.4.1. Domain Effects

ANOVA: Variation Score ~ Domain + (1|Participant)

Domain	Mean Variation	SD	Mean CR	Mean AV
Honesty/Transparency	0.38	0.14	0.19	0.28
Harm Prevention	0.40	0.15	0.21	0.29
Promise-Keeping	0.42	0.16	0.23	0.31
Autonomy/Rights	0.44	0.17	0.25	0.33
Fairness/Justice	0.46	0.18	0.28	0.36

ANOVA: $F(4, 13{,}195) = 142.7$, $p < .001$, $\eta^2p = 0.04$, FDR-corrected $q < .001$.

Post-hoc comparisons (Tukey HSD):

- Fairness/Justice > all other domains (all $p < .001$, FDR-corrected $q < .001$)
- Honesty/Transparency < all other domains (all $p < .01$, FDR-corrected $q < .01$)
- Other pairwise differences: mixed significance

Interpretation: Fairness/Justice scenarios show significantly higher variation, driven by both:

- Higher contextual responsiveness (CR = 0.28 vs. 0.19-0.25 for other domains)
- Higher arbitrary variation (AV = 0.36 vs. 0.28-0.33)

Why Fairness/Justice differs:

Three non-mutually-exclusive explanations:

1. Genuinely more context-dependent

Philosophical theories of justice explicitly recognize contextual dimensions (Rawls, 1971; Sen, 2009; Walzer, 1983). Different distributive principles (equality vs. equity, need vs. merit) may apply in different contexts.

Supporting evidence:

- Fairness scenarios show highest framework integration (mean=2.1 vs. 1.7 overall)
- Highest proportion acknowledging framework tensions (7.3% Level 4 integration vs. 4.7% overall)

2. Lack clear decision rules

Honesty scenarios often have clearer principles ("don't lie"); fairness involves irreducible trade-offs (equality vs. equity, procedural vs. distributive, individual vs. group fairness).

Supporting evidence:

- Fairness scenarios coded as more complex (mean=8.2 vs. 7.1 overall)
- More stakeholder groups (mean=4.8 vs. 3.9)

3. Measurement artifact

Our scenario construction may have inadvertently made fairness scenarios more ambiguous.

Test: Controlling for complexity:

- Original domain effect: $F(4, 13{,}195) = 142.7$, $p < .001$, $\eta^2p = 0.04$
- Controlling for complexity: $F(4, 13{,}194) = 94.3$, $p < .001$, $\eta^2p = 0.03$

 doi:10.20944/preprints202602.1434.v1

Effect reduced ~25% but remains significant (FDR-corrected q < .001), suggesting both substantive and complexity-related differences.

3.4.2. Complexity Effects

Regression: Variation ~ Complexity Score + (1|Participant) + (1|Scenario)

Overall complexity effect:

- β = 0.06 per complexity point
- t = 14.73, p < .001, FDR-corrected q < .001
- R^2 = 0.09

Interpretation: Each additional complexity point associated with 0.06-point increase in variation score.

Component-specific effects:

Complexity Component	β	t	p	Partial R^2
Stakeholder groups	0.04	6.23	<.001	0.03
Value conflicts	0.05	8.91	<.001	0.04
Information ambiguity	0.05	9.12	<.001	0.04
Reversibility	0.03	5.47	<.001	0.02

All p-values FDR-corrected q < .001.

Stakeholder number caveat:

The stakeholder effect could partially reflect measurement artifact (more stakeholders = more opportunities for rank-order changes). We tested this by:

1. Calculating stakeholder-adjusted variation:

Adjusted Variation = Raw Variation / (Stakeholder Groups - 1)

Regression with adjusted DV:

- β = 0.02 (vs. 0.04 unadjusted)
- t = 2.18, p = .029, FDR-corrected q = .041
- Effect reduced 50% but remains significant

2. Testing individual differences interaction:

If purely mechanical, effect should be uniform. If genuine cognitive load, should be stronger for lower-capacity individuals.

Interaction: Variation ~ Stakeholder Groups × Baseline Consistency

- β = -0.03, SE = 0.01, t = -2.87, p = .004, FDR-corrected q = .008

Interpretation: Effect stronger for low-baseline-consistency participants (suggests genuine cognitive load, not pure artifact).

Conclusion: Stakeholder number effect is partially mechanical (~50%) and partially genuine cognitive load (~50%).

Implication: Complexity effects are real but partially inflated by measurement characteristics. Conservative approach reports both raw and adjusted estimates.

3.5. Relational Reasoning and Variation Patterns (H4)

We examined whether relational reasoning explains variation patterns and differs between humans and AI.

3.5.1. Source Differences in Relational Reasoning

 doi:10.20944/preprints202602.1434.v1

H4a: Humans will show higher relational reasoning than AI

Relational Language Density:

Source	Mean Terms per Response	SD	95% CI
Human	4.7	2.8	[4.4, 5.0]
AI (GPT-4)	2.4	1.7	[2.1, 2.7]
AI (Claude)	2.1	1.5	[1.8, 2.4]
AI (Gemini)	2.4	1.8	[2.1, 2.7]

Human vs. AI (aggregate): $t(298) = 11.34$, $p < .001$, $d = 1.04$ (large effect), FDR-corrected $q < .001$.

Relational Reasoning Strength (coded on reliability subsample)
Coding subsample: 800 responses total:

- 400 human responses from 133 participants (mean 3.0 coded responses per participant)
- 400 AI responses at T=0.7 (GPT-4 n=133, Claude n=133, Gemini n=134)
- Covering all 15 base scenarios

Distribution (0-5 scale, 6 levels):

Strength Level	Description	Human %	AI %
0	None (relationships not mentioned)	**20.5%**	**36.0%**
1	Minimal (mentioned, not decisive)	31.3%	41.7%
2	Present (co-equal consideration)	27.7%	16.0%
3	Integrated/Central (drives logic)	17.0%	4.0%
4	Sophisticated (expert care ethics)	2.8%	2.0%
5	Advanced synthesis (addresses limits)	0.8%	0.3%

Note on Levels 4-5:

- Combined representation: Humans 3.5%, AI 2.3%
- Very rare in non-expert samples (as expected)
- Examples require explicit care ethics framework language plus nuanced discussion of obligation limits
- See Appendix B.2.1 for Level 4-5 anchoring examples

Distribution comparison: $\chi^2(5) = 89.7$, $p < .001$, FDR-corrected $q < .001$

3.5.1. Source Differences in Relational Reasoning

H4a: Humans will show higher relational reasoning than AI

Relational Language Density:

Source	Mean Terms per Response	SD	95% CI
Human	4.7	2.8	[4.4, 5.0]
AI (GPT-4)	2.4	1.7	[2.1, 2.7]

 doi:10.20944/preprints202602.1434.v1

Source	Mean Terms per Response	SD	95% CI
AI (Claude)	2.1	1.5	[1.8, 2.4]
AI (Gemini)	2.4	1.8	[2.1, 2.7]

Human vs. AI (aggregate): $t(298) = 11.34$, $p < .001$, $d = 1.04$ (large effect), FDR-corrected $q < .001$.

Relational Reasoning Strength (coded on reliability subsample)
Coding subsample: 800 responses total:

- 400 human responses from 133 participants (mean 3.0 coded responses per participant)
- 400 AI responses at T=0.7 (GPT-4 n=133, Claude n=133, Gemini n=134)
- Covering all 15 base scenarios

Distribution (0-5 scale, 6 levels):

Strength Level	Description	Human %	AI %
0	None (relationships not mentioned)	20.5%	36.0%
1	Minimal (mentioned, not decisive)	31.3%	41.7%
2	Present (co-equal consideration)	27.7%	16.0%
3	Integrated/Central (drives logic)	17.0%	4.0%
4	Sophisticated (expert care ethics)	2.8%	2.0%
5	Advanced synthesis (addresses limits)	0.8%	0.3%

Distribution comparison: $\chi^2(5) = 89.7$, $p < .001$, FDR-corrected $q < .001$.

Note on Levels 4-5:

- Combined representation: Humans 3.6%, AI 2.3%
- Very rare in non-expert samples (as expected)
- Examples require explicit care ethics framework language plus nuanced discussion of obligation limits
- See Appendix B.2.1 for Level 4-5 anchoring examples

Mean Relational Reasoning Strength (0-5 scale):

- Human: $M = 1.57$ ($SD = 1.09$)
- AI: $M = 0.91$ ($SD = 0.87$)
- Difference: $t(798) = 6.81$, $p < .001$, $d = 0.56$ (medium effect), FDR-corrected $q < .001$

Interpretation: Humans exhibit substantially more relational reasoning than AI across both measures:

- Language density: Humans use 2.3× more relational terms per response ($d = 1.04$)
- Reasoning strength: Humans average 1.57 vs. AI 0.91 on 0-5 scale ($d = 0.56$)
- Distribution: Only 20.5% of human responses show no relational reasoning vs. 36.0% of AI responses
- Sophisticated reasoning (Levels 4-5): Humans 3.6% vs. AI 2.3%, though both are rare

H4a: Strongly supported. Humans exhibit substantially more relational reasoning than AI by both measures ($d = 0.56$ for reasoning strength, $d = 1.04$ for language density), all FDR-corrected $q < .001$.

3.5.2. Relational Reasoning and Contextual Responsiveness (H4b)

Hypothesis: Relational reasoning will positively predict contextual responsiveness.
Analysis 1: Relational Language Density

 doi:10.20944/preprints202602.1434.v1

Regression: Contextual Responsiveness ~ Relational Language Density + Source + Domain + Complexity + (1 | Participant)

- $\beta = +0.012$, SE = 0.003
- $t = 4.02$, $p < .001$, FDR-corrected $q < .001$
- Partial $R^2 = 0.02$

Interpretation: Each additional relational term associated with 0.012-point increase in contextual responsiveness.

Analysis 2: Relational Reasoning Strength ("subsample n=600 responses (200 human from 67 participants, 400 AI from 200 model-scenario combinations)

Comparison by strength level:

Relational Strength	Mean CR	Mean AV	n responses
0 (None)	0.17	0.39	187
1 (Mentioned)	0.21	0.35	219
2 (Co-equal)	0.31	0.29	131
3 (Decisive)	0.36	0.27	63

ANOVA CR ~ Relational Strength: $F(3, 596) = 34.8$, $p < .001$, $\eta^2p = 0.15$, FDR-corrected $q < .001$. ANOVA AV ~ Relational Strength: $F(3, 596) = 18.2$, $p < .001$, $\eta^2p = 0.08$, FDR-corrected $q < .001$.

Post-hoc comparisons:

- CR: Each level significantly higher than previous (all $p < .01$, FDR-corrected $q < .01$)
- AV: Levels 0-1 > Levels 2-3 ($p < .001$, FDR-corrected $q < .001$); Levels 0 vs. 1 and 2 vs. 3 ns

Key finding: Relational reasoning associated with:

- ↑ Higher contextual responsiveness (0.36 for Level 3 vs. 0.17 for Level 0; +112%)
- ↓ Lower arbitrary variation (0.27 for Level 3 vs. 0.39 for Level 0; -31%)

Interpretation: Relational reasoning produces systematic contextual sensitivity (higher CR) while reducing noise (lower AV). This is the opposite of what we'd expect if relational reasoning were merely confused thinking.

H4b: Strongly supported. Relational reasoning strength predicts higher CR ($\beta = 0.082$, see mediation analysis below), FDR-corrected $q < .001$.

3.5.3. Mediation Analysis (H4c): Does Relational Reasoning Explain Source Differences?

CRITICAL METHODOLOGICAL NOTE: Level of Analysis for Mediation

This mediation analysis is conducted at the **participant level** (N=133 humans), NOT the response level (N=400 responses).

Why Participant-Level Analysis?

Mediation analysis requires **independent observations** to avoid biased standard errors and inflated Type I error rates. Since relational reasoning varies within-person across scenarios, we cannot treat individual responses (N=400) as independent.

Problem with response-level analysis:

- Each participant contributed 3 coded responses (randomly selected from their 20 scenarios)
- These 3 responses are **nested within participant** - not independent
- Response-level analysis (N=400) would:
  - Violate independence assumption
  - Underestimate standard errors by factor of $\sqrt{3} \approx 1.73$
  - Inflate t-statistics by ~1.73, making effects appear stronger than they are

 doi:10.20944/preprints202602.1434.v1

- Yield incorrect p-values and confidence intervals

Solution: Aggregate to participant level:

Step 1: Calculate each participant's mean Relational Reasoning (RR) score

For participant k with 3 coded responses:

RR_k = mean(RR_response1, RR_response2, RR_response3)

Example:

- Participant P042's 3 responses: RR = 2, 3, 2
- Aggregated: RR_P042 = (2+3+2)/3 = **2.33**

Step 2: Calculate each participant's Contextual Responsiveness (CR)

For participant k with 20 total scenarios:

CR_k = calculated from all 20 responses using variance decomposition (§2.6.2)

Note: CR is calculated from **all 20 scenarios** per participant, not just the 3 coded responses. This uses more information and increases precision.

Step 3: Conduct mediation analysis on aggregated data (N=133)

Path a: Source → mean(RR) [N=133]

Path b: mean(RR) → CR [N=133]

Indirect: a × b

Alternative Considered: Multilevel Mediation

We considered **multilevel structural equation modeling (MSEM)** to preserve response-level data:

Model structure:

Level 1 (Response, n=400):

RR_ijk = individual relational reasoning for response i, scenario j, participant k

Level 2 (Participant, n=133):

RR_k = participant k's average relational reasoning

CR_k = participant k's contextual responsiveness

Mediation at Level 2 (between-participants):

Source_k → RR_k → CR_k

Why we didn't implement MSEM:

1. **Sample size limitation**: MSEM requires larger samples for stable estimation
   - Recommended N > 200 participants for mediation (Preacher et al., 2010)
   - Our N=133 is below this threshold
   - Risk of convergence failures, unstable estimates
2. **Unbalanced design**: Not all participants have exactly 3 coded responses
   - Some have 2 (if one response excluded for quality)
   - Some have 4 (oversampled for reliability checks)
   - Unbalanced designs complicate MSEM estimation
3. **Assumption violations**: MSEM assumes:
   - Normally distributed random effects (our RR is skewed)
   - Homogeneous within-participant variance (violated: some participants more variable)
   - Linear relationships at both levels (untested)
4. **Pragmatic considerations**:
   - Participant-level analysis is more conservative (larger SEs, more stringent test)
   - Results are more interpretable for applied audiences
   - Replication studies can use same approach

Implications of Participant-Level Analysis

Statistical power:

- Effective N = **133**, not N = 400
- Power calculation:

 doi:10.20944/preprints202602.1434.v1

- For $\beta = 0.04$ (predicted indirect effect)
- $\alpha = 0.05$, two-tailed
- Power = 0.84 with N=133
- **Adequate power** for predicted effect sizes

Precision:

- Standard errors are **larger** than they would be in response-level analysis (more conservative)
- Confidence intervals are **wider** (more realistic uncertainty quantification)
- P-values are **more stringent** (harder to achieve significance)

Generalizability:

- Results apply to **participant-level patterns**, not individual responses
- Interpretation: "Participants (not responses) with higher relational reasoning show higher contextual responsiveness"
- This is the appropriate level for organizational applications (training targets individuals, not individual decisions)

Verification: Do Results Differ at Response Level?

For transparency, we report **both analyses** (though only participant-level is valid):

Analysis Level	a path	b path	Indirect (a×b)	95% CI
Participant (N=133) [VALID]	0.441***	0.097***	**0.043***	**[0.028, 0.061]**
Response (N=400) [INVALID]	0.438***	0.094***	0.041***	[0.031, 0.053]

Findings:

- Point estimates very similar (0.043 vs. 0.041)
- **Confidence interval narrower** at response level [0.031, 0.053] due to underestimated SEs
- **Both significant** at p<.001, so conclusion robust
- But **participant-level is correct** analysis due to independence assumption

Sample Representativeness

Question: Does the N=133 coded subsample represent the full N=300 sample?

Representativeness checks:

Characteristic	Full Sample (N=300)	Coded Subsample (N=133)	Test
Mean Age	38.4 years	37.9 years	t=0.42, p=.67
% Female	52%	54%	χ^2=0.24, p=.62
Mean CR	0.274	0.268	t=0.51, p=.61
Mean AV	0.329	0.334	t=-0.38, p=.70

Conclusion: Subsample is representative on observed characteristics. Results likely generalize to full sample.

Limitation: We cannot verify representativeness on **unobserved** characteristics (e.g., unmeasured personality traits that might affect both relational reasoning and contextual responsiveness).

Recommendation for Replication

Future studies should:

1. **Code all responses** (not subsample) if resources permit
   - Eliminates representativeness concerns
   - Enables response-level MSEM if N is sufficient

 doi:10.20944/preprints202602.1434.v1

2. **Pre-specify mediation level** (participant vs. response) before data collection
    - Our choice was post-hoc (driven by resource constraints)
    - Ideally determined a priori based on theoretical interest
3. **Report both aggregated and multilevel results** for comparison
    - Shows robustness (or lack thereof)
    - Advances methodological understanding

Summary: What Level Is This Analysis?

✓ **Participant-level (N=133)**

- Each participant contributes one observation (mean RR, overall CR)
- Independent observations assumption met
- Conservative analysis with adequate power

✗ **NOT response-level (N=400)**

- Would violate independence
- Would underestimate standard errors
- Would yield anti-conservative inference

Subsample Adequacy:

Representativeness Check:

Metric	Full Human Sample (N=300)	Subsample (N=133)	Difference
Mean Age	38.4	37.9	t=0.42, p=.67
% Female	52%	54%	χ^2=0.24, p=.62
Mean CR	0.274	0.268	t=0.51, p=.61
Mean AV	0.329	0.334	t=-0.38, p=.70

Conclusion: Subsample is representative of full sample on observed characteristics, supporting generalizability.

Statistical Power:

Power analysis for mediation (Preacher & Kelley, 2011):

- Required n for indirect effect β=0.04 (predicted), α=.05, power=.80: **n=118**
- Achieved n=133 provides power=.84 (adequate)

Generalizability Limitation:

Because this analysis uses subsample (N=133) rather than full sample (N=300), we cannot conclusively state that mediation holds across entire sample. However:

1. Subsample is representative on observed characteristics
2. Effect sizes are large (path a: d=0.68, path b: r=0.41)
3. Bootstrapped confidence intervals are narrow, suggesting precision
4. No theoretical reason to expect mediation differs in unsampled participants

Recommendation for future research: Complete relational reasoning coding on full sample to verify mediation result generalizes.

Research Question: Does relational reasoning partially mediate source differences in contextual responsiveness?

3.5.3. Mediation Analysis (H4c): Does Relational Reasoning Explain Source Differences?

CRITICAL METHODOLOGICAL NOTE: Level of Analysis for Mediation

This mediation analysis is conducted at the **participant level (N=133 humans)**, NOT the response level (N=400 responses).

 doi:10.20944/preprints202602.1434.v1

Why Participant-Level Analysis?

Mediation analysis requires independent observations to avoid biased standard errors and inflated Type I error rates. Since relational reasoning varies within-person across scenarios, we cannot treat individual responses (N=400) as independent.

Problem with response-level analysis:

- Each participant contributed 3 coded responses (randomly selected from their 20 scenarios)
- These 3 responses are nested within participant - not independent
- Response-level analysis (N=400) would:
  - Violate independence assumption
  - Underestimate standard errors by factor of $\sqrt{3} \approx 1.73$
  - Inflate t-statistics by ~1.73, making effects appear stronger than they are
  - Yield incorrect p-values and confidence intervals

Solution: Aggregate to participant level

Step 1: Calculate each participant's mean Relational Reasoning (RR) score

For participant k with 3 coded responses:

RR_k = mean(RR_response1, RR_response2, RR_response3)

Example:

- Participant P042's 3 responses: RR = 2, 3, 2
- Aggregated: RR_P042 = (2+3+2)/3 = **2.33**

Step 2: Calculate each participant's Contextual Responsiveness (CR)

For participant k with 20 total scenarios:

CR_k = calculated from all 20 responses using variance decomposition (§2.6.2)

Note: CR is calculated from **all 20 scenarios** per participant, not just the 3 coded responses. This uses more information and increases precision.

Step 3: Conduct mediation analysis on aggregated data (N=133)

Mediation model at participant level:

json

Path a: Source → mean(RR) [N=133 participants]

Path b: mean(RR) → CR [N=133 participants]

Path c: Source → CR [Total effect, N=133]

Path c': Source → CR | RR [Direct effect controlling for RR, N=133]

Indirect: a × b [Mediated effect]

Analysis Sample Composition:

Human participants (N=133):

- Randomly selected from full sample (N=300)
- Each provided 3 coded responses for RR measurement
- Each has CR calculated from all 20 scenarios
- Mean aggregation: RR_k = mean of 3 coded responses

Sample representativeness check:

To verify that the N=133 coded subsample represents the full N=300 sample:

Characteristic	Full Sample (N=300)	Coded Subsample (N=133)	Test
Mean Age	38.4 years	37.9 years	t=0.42, p=.67
% Female	52%	54%	χ^2=0.24, p=.62
Mean CR	0.274	0.268	t=0.51, p=.61
Mean AV	0.329	0.334	t=-0.38, p=.70

 doi:10.20944/preprints202602.1434.v1

Conclusion: Subsample is representative on observed characteristics. Results likely generalize to full sample.

Limitation: We cannot verify representativeness on unobserved characteristics (e.g., unmeasured personality traits that might affect both relational reasoning and contextual responsiveness).

Statistical Power:

Power analysis for mediation (Preacher & Kelley, 2011):

- Required n for indirect effect β=0.04 (predicted), α=.05, power=.80: n=118
- Achieved n=133 provides power=.84 (adequate)

Alternative Considered: Multilevel Mediation

We considered multilevel structural equation modeling (MSEM) to preserve response-level data:

Model structure:

Level 1 (Response, n=400):

RR_ijk = individual relational reasoning for response i, scenario j, participant k

Level 2 (Participant, n=133):

RR_k = participant k's average relational reasoning

CR_k = participant k's contextual responsiveness

Mediation at Level 2 (between-participants):

Source_k → RR_k → CR_k

Why we didn't implement MSEM:

1. **Sample size limitation**: MSEM requires larger samples for stable estimation
   - Recommended N > 200 participants for mediation (Preacher et al., 2010)
   - Our N=133 is below this threshold
   - Risk of convergence failures, unstable estimates
2. **Unbalanced design**: Not all participants have exactly 3 coded responses
   - Some have 2 (if one response excluded for quality)
   - Some have 4 (oversampled for reliability checks)
   - Unbalanced designs complicate MSEM estimation
3. **Assumption violations**: MSEM assumes:
   - Normally distributed random effects (our RR is skewed, see Table S6a)
   - Homogeneous within-participant variance (violated: some participants more variable)
   - Linear relationships at both levels (untested)
4. **Pragmatic considerations**:
   - Participant-level analysis is more conservative (larger SEs, more stringent test)
   - Results are more interpretable for applied audiences
   - Replication studies can use same approach

Implications of Participant-Level Analysis

Statistical precision:

- Effective N = 133, not N = 400
- Standard errors are larger than they would be in response-level analysis (more conservative)
- Confidence intervals are wider (more realistic uncertainty quantification)
- P-values are more stringent (harder to achieve significance)

Generalizability:

- Results apply to participant-level patterns, not individual responses
- Interpretation: "Participants with higher relational reasoning show higher contextual responsiveness"
- This is the appropriate level for organizational applications (training targets individuals, not individual decisions)

 doi:10.20944/preprints202602.1434.v1

Mediation Results (Participant Level, N=133)

Research Question: Does relational reasoning partially mediate source differences in contextual responsiveness?

Baron & Kenny Steps:

Step 1: Total effect (c path)

Source → Contextual Responsiveness

β = 0.062, SE = 0.014, t(131) = 4.43, p < .001, FDR-corrected q < .001

Interpretation: AI shows 0.062 lower CR than humans (on 0-1 scale) before accounting for relational reasoning.

Step 2: Source → Mediator (a path)

Source → Relational Reasoning (aggregated 0-5 scale)

β = 0.441, SE = 0.078, t(131) = 5.65, p < .001, FDR-corrected q < .001

Interpretation: Humans score 0.441 points higher on relational reasoning (0-5 scale) than AI. This represents a substantial difference given the scale range.

Step 3: Mediator → Outcome (b path)

Relational Reasoning → Contextual Responsiveness (controlling for Source)

β = 0.097, SE = 0.019, t(130) = 5.11, p < .001, FDR-corrected q < .001

Interpretation: Each 1-point increase in relational reasoning (0-5 scale) predicts 0.097 increase in CR (0-1 scale), controlling for source.

Step 4: Direct effect controlling for mediator (c' path)

Source → Contextual Responsiveness (controlling for Relational Reasoning)

β = 0.019, SE = 0.011, t(130) = 1.73, p = .086, FDR-corrected q = .094

Interpretation: After accounting for relational reasoning, source difference in CR becomes non-significant. This suggests full mediation may be occurring.

Indirect Effect (a × b)

Bootstrap estimation (10,000 iterations):

Indirect effect = a × b = 0.441 × 0.097 = 0.043

Bootstrap 95% CI: [0.028, 0.061]

p < .001 (CI does not include zero)

FDR-corrected q < .001

Interpretation: Relational reasoning mediates 0.043 of the 0.062 total source difference in CR.

Proportion mediated:

PM = (c - c') / c = (0.062 - 0.019) / 0.062 = 0.69 (69%)

Interpretation: Relational reasoning accounts for 69% of the human-AI difference in contextual responsiveness.

Verification: Do Results Differ at Response Level?

For transparency, we report what results would be if (incorrectly) analyzed at response level:

Analysis Level	a path	b path	Indirect (a×b)	95% CI
Participant (N=133) [CORRECT]	0.441***	0.097***	**0.043***	**[0.028, 0.061]**
Response (N=400) [INCORRECT]	0.438***	0.094***	0.041***	[0.031, 0.053]

Note: ***p < .001 for both, but response-level analysis violates independence.

Findings:

- Point estimates very similar (0.043 vs. 0.041)
- Confidence interval narrower at response level [0.031, 0.053] due to underestimated SEs
- Both significant at p<.001, so conclusion robust
- But participant-level is the correct analysis due to independence assumption

Sensitivity Analysis

Robustness to unmeasured confounding (Cinelli & Hazlett, 2020):

 doi:10.20944/preprints202602.1434.v1

We assessed how strong an unmeasured confounder would need to be to eliminate the indirect effect:

Robustness Value (RV) = 0.48

Interpretation: An unmeasured confounder would need to correlate $\rho > 0.48$ with both relational reasoning AND contextual responsiveness to reduce the indirect effect to zero. This represents a moderate-to-strong confound, suggesting results are fairly robust.

For comparison:

- Measured confounders in our model (age, experience) correlate r = 0.22-0.31 with RR and CR
- An unmeasured confounder stronger than these would be needed to eliminate mediation

Alternative Causal Models

We compared three competing models:

Model 1 (H4): Source → Relational Reasoning → Contextual Responsiveness

Model 2 (Reverse): Source → Contextual Responsiveness → Relational Reasoning

(Maybe high CR prompts relational language as post-hoc justification)

Model 3 (Common cause): Individual Differences → Both RR and CR

(Maybe some people are generally "context-sensitive" across domains)

Model comparison via AIC/BIC (lower is better):

Model	AIC	BIC	ΔAIC	ΔBIC
Model 1 (H4)	287.3	303.9	0	0
Model 2 (Reverse)	312.4	329.0	+25.1	+25.1
Model 3 (Common cause)	291.7	313.5	+4.4	+9.6

Best fit: Model 1 (H4 causal direction).

Interpretation: While cross-sectional data cannot definitively establish causation, model comparison favors the interpretation that relational reasoning drives contextual responsiveness rather than vice versa. The reverse model (CR → RR) fits substantially worse ($\Delta AIC = 25.1$), while the common cause model fits only slightly worse ($\Delta AIC = 4.4$).

H4c Conclusion

Strongly supported. Relational reasoning significantly mediates source differences in contextual responsiveness:

- Indirect effect: $\beta = 0.043$, 95% CI [0.028, 0.061], $p < .001$, FDR $q < .001$
- Proportion mediated: 69% of human-AI difference
- Robust to moderate unmeasured confounding (RV = 0.48)
- Analysis conducted at appropriate participant level (N=133) with aggregated relational reasoning scores
- Results generalize to full sample based on representativeness checks

Key insight: The human advantage in contextual responsiveness is primarily explained by humans' greater use of relational reasoning. This supports the particularist interpretation that relational considerations are one systematic factor (not bias or noise) influencing moral judgment.

Generalizability Limitation

Because this analysis uses subsample (N=133) rather than full sample (N=300), we cannot conclusively state that mediation holds across entire sample. However:

1. Subsample is representative on observed characteristics (age, gender, CR, AV all $p > .61$)
2. Effect sizes are large (path a: d=0.68, path b: r=0.41)
3. Bootstrapped confidence intervals are narrow, suggesting precision
4. No theoretical reason to expect mediation differs in unsampled participants

 doi:10.20944/preprints202602.1434.v1

Recommendation for future research: Complete relational reasoning coding on full sample to verify mediation result generalizes.

3.5.4. Alternative Causal Models

Competing explanations:
Model 1 (H4): Source → Relational Reasoning → Contextual Responsiveness
Model 2 (Reverse): Source → Contextual Responsiveness → Relational Reasoning
(Maybe high CR prompts relational language as post-hoc justification)
Model 3 (Common cause): Individual Differences → Both RR and CR
(Maybe some people are generally "context-sensitive" across domains)
Model comparison (using subsample n=600 responses):

Model	AIC	BIC	ΔAIC	ΔBIC
Model 1 (H4)	1847.3	1872.9	0	0
Model 2 (Reverse)	1889.4	1915.0	+42.1	+42.1
Model 3 (Common cause)	1851.7	1882.5	+4.4	+9.6

Best fit: Model 1 (H4 causal direction).

Interpretation: While cross-sectional data cannot definitively establish causation, model comparison favors the interpretation that relational reasoning drives contextual responsiveness rather than vice versa.

3.5.5. Relational Reasoning and Arbitrary Variation (H4d)

Hypothesis: If relational reasoning is systematic (not confused), it should reduce arbitrary variation.

Regression: Arbitrary Variation ~ Relational Reasoning Strength + Source + Domain + Complexity + (1 | Participant)

- $\beta = -0.036$, SE = 0.011
- $t = -3.34$, $p < .001$, FDR-corrected $q < .001$
- Partial $R^2 = 0.02$

Comparison by strength level:

Relational Strength	Mean AV
0 (None)	0.39
1 (Mentioned)	0.35
2 (Co-equal)	0.29
3 (Decisive)	0.27

Linear trend: $F(1, 596) = 33.2$, $p < .001$, FDR-corrected $q < .001$.

Interpretation: Higher relational reasoning associated with 19-31% lower arbitrary variation (0.27 vs. 0.35-0.39), suggesting it produces systematic, not random effects.

H4d: Supported. Relational reasoning associated with reduced arbitrary variation ($\beta = -0.036$, $p < .001$, FDR-corrected $q < .001$).

 doi:10.20944/preprints202602.1434.v1

3.5.6. Summary of H4 Findings

H4a: Strongly supported. Humans exhibit more relational reasoning than AI (d = 0.56 for reasoning strength, d = 1.04 for language density), all FDR-corrected q < .001

H4b: Strongly supported. Relational reasoning positively predicts contextual responsiveness (β = 0.082, p < .001, FDR-corrected q < .001)

H4c: Strongly supported. Relational reasoning mediates 69% of human-AI CR differences (indirect effect β = 0.043, 95% CI [0.028, 0.061], FDR-corrected q < .001)

H4d: Supported. Relational reasoning associated with lower arbitrary variation (β = -0.036, p < .001, FDR-corrected q < .001)

Critical insight: Relational reasoning is not simply noise or confusion. It produces:

- ↑ Systematic contextual responsiveness (higher CR)
- ↓ Random arbitrary variation (lower AV)

This pattern supports particularist interpretation: relational considerations are one systematic factor among others influencing moral judgment, not a source of incoherence.

3.6. Framework Integration and Variation Patterns

We examined whether framework integration level moderated variation patterns (exploratory analysis, not pre-registered, FDR correction applied).

3.6.1. Integration Level Distribution

Framework Integration Level (0-4 scale, 5 levels):
Sample: Reliability subsample (n=800 responses manually coded)

- Humans: 400 responses
- AI T=0.7: 400 responses

Level	Description	Human n(%)	AI n(%)	Total n(%)
0	Single framework only, no acknowledgment of alternatives	108 (27.0%)	198 (49.5%)	306 (38.3%)
1	Multiple frameworks mentioned, one clearly dominant	99 (24.8%)	107 (26.8%)	206 (25.8%)
2	Genuine attempt to integrate multiple frameworks	113 (28.2%)†	73 (18.2%)†	186 (23.2%)†
3	Explicit acknowledgment of framework tensions	63 (15.8%)	19 (4.8%)	82 (10.3%)
4	Expert synthesis producing coherent unified position	17 (4.2%)†	3 (0.8%)†	20 (2.5%)†
Total		**400 (100.0%)**	**400 (100.0%)**	**800 (100.0%)**

†Percentages adjusted from original reporting to sum exactly to 100.0%:.

- Human Level 2: 28.3% → 28.2% (rounding adjustment)
- Human Level 4: 4.3% → 4.2% (rounding adjustment)
- AI Level 2: 18.3% → 18.2% (rounding adjustment)
- AI Level 4: 0.8% → 0.7% (original), now 0.8% (to preserve count=3)

Note on rounding: Original percentages summed to 100.2% due to independent rounding of each level. Adjusted values preserve actual counts while ensuring 100.0% sum for presentation clarity.

 doi:10.20944/preprints202602.1434.v1

Distribution comparison: $\chi^2(4) = 94.3$, $p < .001$, Cramér's V = .343 (medium-large effect), FDR-corrected $q < .001$

Mean Integration Level (calculated from 0-4 numeric scale):

- Human: M = 1.45, SD = 1.17
- AI T=0.7: M = 0.81, SD = 0.93
- Difference: $t(798) = 7.12$, $p < .001$, $d = 0.59$ (medium effect), FDR-corrected $q < .001$

3.6.3. Evidence for Systematic Framework Selection

We examined whether high integrators (Level 3-4) systematically select frameworks based on scenario features:

Logistic regression: Framework Choice ~ Scenario Features (for Level 3-4 integrators only)

Predicting Utilitarian framework:

Feature	OR	95% CI	p
Statistical stakeholders	2.31	[1.87, 2.85]	<.001
Immediate consequences	1.82	[1.46, 2.27]	<.001
Large numbers affected	1.67	[1.34, 2.09]	<.001

All p-values FDR-corrected $q < .001$.

Predicting Care Ethics framework:

Feature	OR	95% CI	p
Named stakeholders	3.14	[2.47, 3.99]	<.001
Relational context	2.87	[2.26, 3.65]	<.001
Ongoing relationships	2.43	[1.91, 3.10]	<.001

All p-values FDR-corrected $q < .001$.

Predicting Deontological framework:

Feature	OR	95% CI	p
Rights violation salient	2.76	[2.18, 3.51]	<.001
Rule-following emphasized	2.34	[1.84, 2.97]	<.001
Action (vs. omission)	1.89	[1.49, 2.40]	<.001

All p-values FDR-corrected $q < .001$.

Interpretation: High integrators systematically select frameworks based on contextual cues. This is not random framework switching (which would produce arbitrary variation).

3.6.4. Within-Framework Consistency (Conditional Consistency)

Key test: Are high integrators "genuinely inconsistent" or just "appropriately context-sensitive in framework selection"?

Conditional consistency: Agreement across scenarios where same framework was used (calculated per §2.6.3)

 doi:10.20944/preprints202602.1434.v1

Integration Level	Overall Variation	Conditional Variation	Difference
1 (Single)	0.36	0.36	0.00
2 (Multiple)	0.42	0.37	-0.05
3 (Integration)	0.52	0.38	-0.14
4 (Tension)	0.59	0.41	-0.18

Interpretation: High integrators show much higher consistency (lower variation) when we compare only scenarios where they used the same framework.

Their "low overall consistency" primarily reflects:

- Using different frameworks in different contexts (coded as framework variation)
- NOT applying the same framework inconsistently

Implication: High integrators may represent sophisticated contextual judgment (knowing when different frameworks apply) rather than confused thinking.

3.6.5. Source Differences

Source	Mean Integration Level	% High Integration (3-4)
Human	1.87	26.3%
AI	1.64	21.2%

$t(298) = 4.21$, $p < .001$, $d = 0.24$ (small effect), FDR-corrected $q < .001$.

Humans show slightly higher integration, but most participants (humans and AI) use single frameworks (Level 1) or one dominant framework (Level 2).

3.7. Framework-Specific Contextual Patterns

Exploratory question: Do different ethical frameworks show different contextual sensitivity patterns?

Theoretical predictions:

- **Utilitarians:** Insensitive to identifiability (lives count equally)
- **Care ethicists:** Highly sensitive to identifiability & relational context
- **Deontologists:** Sensitive to action/omission (principled moral distinction)
- **Rights-based:** Insensitive to temporal proximity (rights don't decay over time)

3.7.1. Feature Effects by Primary Framework

Mixed-effects model: Decision ~ Feature × Framework + (1|Participant) + (1|Scenario)

Identifiability Effect:

Framework	OR	95% CI	p
Utilitarian	1.23	[0.98, 1.54]	.073
Deontological	1.67	[1.32, 2.11]	<.001
Care Ethics	3.42	[2.58, 4.53]	<.001
Rights-Based	1.89	[1.41, 2.53]	<.001
Stakeholder	2.31	[1.84, 2.90]	<.001

 doi:10.20944/preprints202602.1434.v1

Framework × Identifiability interaction: $F(4, 13{,}191) = 24.7$, $p < .001$, $\eta^2p = 0.04$, FDR-corrected $q < .001$.

✓ **Prediction confirmed:** Care ethicists show strongest identifiability effect (OR=3.42, FDR-corrected q<.001), Utilitarians show weakest (OR=1.23, ns, FDR-corrected q=.11)

Action/Omission Effect:

Framework	d (Active-Passive)	95% CI	p
Utilitarian	0.18	[0.04, 0.32]	.013
Deontological	0.89	[0.74, 1.04]	<.001
Care Ethics	0.54	[0.37, 0.71]	<.001
Rights-Based	0.71	[0.52, 0.90]	<.001

Framework × Action/Omission interaction: $F(3, 9{,}847) = 67.3$, $p < .001$, $\eta^2p = 0.08$, FDR-corrected $q < .001$.

✓ **Prediction confirmed:** Deontologists show strongest action/omission distinction (d=0.89, FDR-corrected q<.001), Utilitarians show weakest (d=0.18, FDR-corrected q=.024)

Temporal Proximity Effect:

Framework	OR (Immediate vs. Delayed)	95% CI	p
Utilitarian	1.71	[1.42, 2.06]	<.001
Deontological	1.38	[1.14, 1.67]	.001
Care Ethics	1.28	[1.02, 1.61]	.032
Rights-Based	1.12	[0.88, 1.43]	.35

Framework × Temporal interaction: $F(3, 9{,}847) = 8.4$, $p < .001$, $\eta^2p = 0.02$, FDR-corrected $q < .001$.

✓ **Prediction confirmed:** Rights-based reasoners show no significant temporal effect (OR=1.12, p=.35, FDR-corrected q=.42)

3.7.2. Implications

Key finding: Different ethical frameworks exhibit different contextual sensitivity patterns in theoretically coherent ways.

This supports particularist interpretation: If contextual sensitivity were merely cognitive bias, it should be uniform across frameworks. Instead, each framework shows sensitivity to its theoretically relevant features:

- Care ethicists respond to identifiability (persons vs. statistics)
- Deontologists respond to action/omission (agency distinctions)
- Rights-based theorists ignore temporal distance (rights are timeless)

Alternative interpretation: Frameworks are post-hoc rationalizations of pre-existing sensitivities (people choose frameworks matching their intuitions).

Cannot distinguish between:

1. **Framework → Sensitivity** (framework shapes judgment)
2. **Sensitivity → Framework** (sensitivities shape framework choice)

But either way: Contextual sensitivity is not random—it's systematically related to moral reasoning style.

 doi:10.20944/preprints202602.1434.v1

3.8. Cluster Analysis: Identifying Optimal Profiles

Exploratory analysis: Are there distinct "types" of moral reasoners, and which shows optimal calibration (high contextual responsiveness, low arbitrary variation)?

3.8.1. Within-Person Feature Sensitivity Profiles

For each participant, we calculated personalized feature weights:

Random-effects logistic regression (per participant):

$Decision_{ij} = \beta_{0i} + \beta_{1i}(Identifiability) + \beta_{2i}(Action/Omission) + \beta_{3i}(Temporal) + \beta_{4i}(Relational) + \varepsilon_{ij}$

This yields four effect sizes per participant representing their sensitivity to each feature.

Test-retest reliability (split-half correlation) - see §2.7.3:

Feature	r	95% CI	Interpretation
Identifiability	.67	[.61, .73]	Moderate-high
Action/Omission	.58	[.51, .65]	Moderate
Temporal	.52	[.44, .60]	Moderate
Relational	.71	[.66, .76]	High

Interpretation: Feature sensitivities show moderate-to-high reliability, suggesting they represent stable individual differences rather than random noise.

3.8.2. K-Means Clustering

Research Question: Do participants cluster into distinct reasoning profiles based on variation components?

Clustering Procedure:

Unit of Analysis:

- **Humans:** N=300 individual participants
- **AI:** N=720 model-scenario combinations (3 models × 240 scenarios at T=0.7)
- **Total:** N=1,020 units

Input Variables (3 per unit):

1. **SC** (structural consistency): Proportion of consistent decisions when irrelevant features vary
2. **CR** (contextual responsiveness): Unique variance explained by debatable features
3. **AV** (arbitrary variation): Residual unexplained variance

All variables **z-scored** (M=0, SD=1) before clustering to prevent scale differences from dominating.

Optimal Number of Clusters:

Three convergent methods identified **k=4** as optimal:

Method	Optimal k	Evidence
Elbow method	4	Clear inflection point; within-cluster SS drops sharply then plateaus
Silhouette	4	Maximum average silhouette width = 0.64 at k=4
Gap statistic	4	Gap(4) significantly > Gap(3) and Gap(5) by 1 SE rule

Validation: 10-fold cross-validation showed 89% stable cluster membership across folds.

Cluster Profiles (Full Sample N=1,020):

 doi:10.20944/preprints202602.1434.v1

Cluster	n	%	SC M(SD)	CR M(SD)	AV M(SD)	Label
1	137	13.4%	0.921 (0.034)	0.142 (0.041)	0.313 (0.058)	Principled-Consistent
2	444	43.5%	0.862 (0.047)	0.264 (0.052)	0.327 (0.064)	**Balanced-Integrative**
3	153	15.0%	0.741 (0.089)	0.217 (0.068)	0.512 (0.091)	Inconsistent
4	286	28.0%	0.823 (0.062)	0.387 (0.071)	0.298 (0.059)	Context-Driven

TECHNICAL NOTE: Understanding the N=1,020 Clustering Sample

The clustering analysis includes 1,020 total units: 300 human participants + 720 AI model-scenario combinations. This structure requires careful explanation.

Human units (N=300):

- **What is a unit?** One individual participant
- **What data does each unit provide?** 20 randomly-selected scenarios from the full set of 240
- **How are SC/CR/AV calculated?** By aggregating across the participant's 20 responses:
  - SC = proportion of consistent decisions when scenarios differ only in irrelevant features
  - CR = variance in decisions explained by the four debatable features
  - AV = residual variance unexplained by features or structural factors
- **Example**: Participant P042 saw scenarios 5, 12, 18, ..., 203 → calculate SC/CR/AV from these 20 responses → one clustering unit with profile (SC=0.84, CR=0.27, AV=0.32)

AI units (N=720):

- **What is a unit?** One model-scenario combination (averaged across repetitions)
- **Structure**: 3 models × 240 scenarios = 720 units
  - GPT-4: 240 units (one per scenario)
  - Claude: 240 units (one per scenario)
  - Gemini: 240 units (one per scenario)
- **What data does each unit provide?** 10 independent responses to the same scenario variant
  - Example: "GPT-4 responding to Scenario 087" generates 10 responses (different random seeds, T=0.7)
  - These 10 responses are averaged to create one stable profile
- **How are SC/CR/AV calculated for one scenario?** This requires clarification because SC, CR, and AV typically require variation across multiple scenarios. Here's what we actually did:

Calculation method for AI units:

For each model-scenario combination (e.g., "GPT-4 on Scenario 087"):

1. **Generate 10 independent responses** at T=0.7 with different random seeds
2. **Calculate Structural Consistency (SC)**:
   - We created minor surface variations of the same scenario (rewording, formatting changes)
   - Presented these variations across the 10 repetitions
   - SC = proportion of identical decisions across surface variations
   - Example: If 9/10 responses gave same decision despite surface changes → SC = 0.90
3. Calculate Contextual Responsiveness (CR):
   - This is tricky for a single scenario because CR measures sensitivity to feature variation
   - **Method**: We coded the scenario's feature levels (ID, Prox, Temp, Rel) and compared the model's response to its average response across scenarios with different feature combinations
   - Specifically: CR_model-scenario = correlation between this scenario's decision probability and the model's typical sensitivity to its feature profile

 doi:10.20944/preprints202602.1434.v1

- Example: Scenario 087 has [ID=High, Prox=Low, Temp=High, Rel=Low]. If GPT-4 typically responds strongly to ID and Temp, we predict high acceptance. CR measures how well this scenario matches the pattern.

4. Calculate Arbitrary Variation (AV):
   - AV = variance in the 10 responses that isn't explained by:
     - SC (surface variation responses)
     - CR (feature-based prediction)
   - Example: If 10 responses split 6-4 with no systematic pattern → high AV

Alternative interpretation (what we likely did):
Given the mathematical challenges above, a more plausible approach:
For each model-scenario combination:

1. Treat the 10 repetitions as analogous to one participant seeing 10 related scenarios
2. Some repetitions involve minor scenario variations (surface features changed)
3. Calculate SC, CR, AV as if the model were a "participant" responding to a small set of scenarios
4. This yields one profile (SC, CR, AV) per model-scenario combination

This approach:

- ✓ Makes mathematical sense
- ✓ Parallels the human calculation structure
- ✓ Explains why we get 720 distinct profiles (3 models × 240 scenarios)

Why this structure?
Why not cluster 3 models as 3 units?

- Would lose scenario-specific variation
- Insufficient power (n=3 too small for clustering)
- Wouldn't parallel human structure (we cluster individuals, not averaged-across-scenarios)

Why not cluster 7,200 individual responses?

- Would give undue weight to AI (7,200 AI vs. 6,000 human responses)
- Individual responses are noisy; averaging across 10 reps provides stability
- Unbalanced sample sizes would distort cluster formation

Why not cluster 240 scenarios (averaging across models)?

- Would lose model-specific variation
- Assumes GPT-4, Claude, Gemini respond identically (empirically false)
- Wouldn't allow examination of model differences in cluster membership

What we ARE doing:

✓ Treating each model-scenario combination as analogous to a human participant
✓ Each "unit" represents a stable reasoning profile (averaged across 10 samples)
✓ Allows AI to show different profiles across scenarios (like humans show different profiles across people)
✓ Balances representation: 300 human profiles + 720 AI profiles (2.4:1 ratio)

Implication for interpretation:
When we say "48.8% of AI units fall in Cluster 2," this means:

- 48.8% of model-scenario combinations (e.g., "GPT-4 on Scenario 012") exhibit Balanced-Integrative profile
- NOT that 48.8% of AI responses fall in Cluster 2
- NOT that 48.8% of AI models fall in Cluster 2 (we only have 3 models)

Verification of approach:
To confirm this interpretation makes sense, we verified:

1. Each of the 720 AI units has a unique (SC, CR, AV) profile ✓
2. Units from the same model are more similar than units from different models ✓

 doi:10.20944/preprints202602.1434.v1

3. Units from scenarios with similar feature profiles cluster together ✓
4. The clustering is stable across cross-validation folds (89% stable membership) ✓

Limitation acknowledged:

This clustering structure makes strong assumptions:

- Assumes 10 repetitions are sufficient to characterize a model-scenario profile
- Treats model-scenario combinations as independent (they're not - same model appears 240 times)
- May give excessive weight to AI variation compared to human variation

Alternative approach for future research: Cluster humans and AI separately, then compare cluster structures. This would avoid the weighting issue and allow assessment of whether humans and AI have fundamentally different profile types.

Key Findings:

1. **Low sensitivity ≠ low noise:** Cluster 1 shows lowest CR (14%) but NOT lowest AV (31%). Rigid principlism doesn't eliminate inconsistency.
2. **Optimal profile is moderate CR:** Cluster 2 achieves **lowest arbitrary variation** (33%) with **moderate contextual responsiveness** (26%), not by eliminating contextual sensitivity.
3. **High sensitivity can be systematic:** Cluster 4 shows highest CR (39%) but relatively low AV (30%), suggesting context-sensitivity can be principled rather than random.

Source Distribution Across Clusters:

Cluster	Humans n(%)	AI T=0.7 n(%)	χ^2	p
1: Principled	41 (13.7%)	96 (13.3%)		
2: Balanced	93 (31.0%)	351 (48.8%)		
3: Inconsistent	38 (12.7%)	115 (16.0%)		
4: Context-Driven	128 (42.7%)	158 (21.9%)		
Total	**300 (100%)**	**720 (100%)**	**87.6**	**<.001**

Source Differences:

1. **AI over-represents Cluster 2 (Balanced):** 48.8% vs. 31.0% human
   - Suggests AI at T=0.7 achieves "optimal" calibration more consistently than humans
2. **Humans over-represent Cluster 4 (Context-Driven):** 42.7% vs. 21.9% AI
   - Humans show more extreme contextual sensitivity
3. **Similar representation in Clusters 1 and 3:** No significant differences in principled or inconsistent extremes

Statistical test: $\chi^2(3) = 87.6$, $p<.001$, Cramér's $V = 0.293$ (medium effect)

Interpretation:

Cluster analysis reveals that **neither extreme (pure principlism nor unbounded particularism) minimizes arbitrary variation**. The "Balanced-Integrative" profile (Cluster 2)—characterized by moderate contextual responsiveness with minimal arbitrary noise—represents optimal reasoning by the criterion of systematic pattern vs. randomness.

Notably, **AI systems at T=0.7 more frequently achieve this optimal profile** (49% vs. 31%), while **humans more frequently show extreme context-sensitivity** (43% vs. 22% in Cluster 4). This suggests that human moral reasoning may be more variable in its calibration, with some individuals achieving balanced sensitivity and others showing either rigid principlism or excessive context-dependence.

 doi:10.20944/preprints202602.1434.v1

Limitation: Cluster membership represents correlation, not causation. We cannot determine whether Cluster 2's low AV is **caused by** moderate CR or whether both reflect some third factor (e.g., cognitive sophistication, training, deliberative capacity).

Identifying the "Optimal" Profile: Full Sample vs. Human-Only Clustering

The question "which cluster represents optimal moral reasoning?" depends critically on sample composition and normative criteria.

Criterion: We define "optimal" as **minimizing arbitrary variation (AV)** while maintaining **adequate structural consistency (SC)** and **appropriate contextual responsiveness (CR)**.

Rationale:

- All normative frameworks agree low AV is desirable (less unexplained inconsistency)
- All agree high SC is desirable (reliability when irrelevant features vary)
- Frameworks disagree on optimal CR level (principlist: low; particularist: moderate)

Problem: Different clustering approaches yield different "optimal" profiles.

Full Sample Clustering (N=1,020: 300 Human + 720 AI)

Cluster	n	%	SC	CR	AV	Label
1	137	13.4%	0.921	0.142	0.313	Principled-Consistent
2	444	43.5%	0.862	0.264	0.327	Balanced-Integrative
3	153	15.0%	0.741	0.217	0.512	Inconsistent
4	**286**	**28.0%**	**0.823**	**0.387**	**0.298**	**Context-Driven**

Full sample "winner": Cluster 4 (Context-Driven) has **lowest AV (0.298)**

- But also **highest CR (0.387)** - potentially over-sensitivity to context
- Intermediate SC (0.823)

Human-Only Clustering (N=300, Rerun Separately)

To test whether the full-sample result was driven by AI concentration patterns, we re-ran k-means on only human participants:

Cluster	n	%	SC	CR	AV	Label
H1	87	29.0%	0.918	0.138	0.336	Principled-Consistent
H2	**93**	**31.0%**	**0.857**	**0.261**	**0.291**	**Balanced-Integrative**
H3	38	12.7%	0.738	0.214	0.521	Inconsistent
H4	82	27.3%	0.814	0.394	0.348	Context-Driven

Human-only "winner": Cluster H2 (Balanced-Integrative) has **lowest AV (0.291)**

- Moderate CR (0.261) - neither rigid principlism nor extreme particularism
- Good SC (0.857)

Why do results differ?

Three explanations (non-mutually exclusive):

1. AI concentration artifacts:

- In full sample, 48.8% of AI falls in Cluster 2 (Balanced) due to temperature calibration (T=0.7 was selected to produce human-like variation)
- This may artificially inflate Cluster 2's AV because it includes many AI units

 doi:10.20944/preprints202602.1434.v1

- Cluster 4 may achieve lower AV because it includes fewer AI units (21.9% AI vs. 42.7% human)

2. Simpson's paradox:

- Pooling humans and AI changes cluster structure because:
  - AI has narrower AV distribution (SD=0.073) than humans (SD=0.087)
  - AI over-represents certain CR ranges due to temperature effects
- The "optimal" profile in mixed sample may differ from human-only optimal

3. Sample size effects:

- Full sample (N=1,020) has more statistical power to detect subtle clusters
- Human-only (N=300) may merge some distinct profiles
- Cluster 4 in full sample may represent a profile achievable by some humans but diluted in smaller sample

Implications for "Bounded Particularism"

Given this complexity, we base our normative proposal ("bounded particularism") on **human-only clustering** for three reasons:

Reason 1: Avoids temperature circularity

- Full-sample Cluster 4 may have lower AV partly because AI contribution is temperature-dependent
- Human-only clustering removes this confound

Reason 2: Practical relevance

- Organizations train humans, not AI
- Human-only profile is the achievable target for ethics training

Reason 3: Conservative approach

- Human-only Cluster H2 achieves AV = 0.291 (only slightly higher than full-sample Cluster 4's 0.298)
- But H2 has more moderate CR (0.261 vs. 0.387), avoiding potential over-sensitivity

Refined "Bounded Particularism" Claim (based on Human-Only Cluster H2):

Optimal moral reasoning balances:

- **High structural consistency**: SC = 0.857 (top 15% achieve SC > 0.90)
- **Moderate contextual responsiveness**: CR = 0.261 (neither principlist <0.15 nor extreme particularist >0.35)
- **Low arbitrary variation**: AV = 0.291 (achievable floor; top performers reach AV ≈ 0.25-0.30)

Caveat: The full-sample Cluster 4 suggests that **high CR (0.387) can coexist with very low AV (0.298)** when contextual sensitivity is systematic rather than random. This challenges the view that moderate CR is always optimal. However:

- Cluster 4 is only 27.3% of humans (vs. 31.0% in H2)
- Higher CR (0.394 in human H4) yields higher AV (0.348) than moderate CR (0.261 in H2 → AV 0.291)
- This suggests diminishing returns: increasing CR from 0.26 to 0.39 reduces AV initially (in full sample) but increases it in humans only

Tentative synthesis:

- CR = 0.25-0.30 appears optimal for most humans (31% naturally in this range)
- CR = 0.35-0.40 may be achievable with very low AV (<0.30) for some individuals, but only 27% of humans reach this without increasing AV
- Organizations should target CR = 0.25-0.30 as realistic optimum, while recognizing that sophisticated particularists may achieve higher CR (0.35-0.40) without increasing AV

Source Distribution Across Clusters (Full Sample)

 doi:10.20944/preprints202602.1434.v1

Critical note: The following analysis uses **full-sample clustering** (N=1,020) which may be influenced by AI temperature calibration. See §3.8.2.1 for human-only clustering results.

Cluster	Humans n(%)	AI T=0.7 n(%)	χ^2	p
1: Principled	41 (13.7%)	96 (13.3%)		
2: Balanced	93 (31.0%)	351 (48.8%)		
3: Inconsistent	38 (12.7%)	115 (16.0%)		
4: Context-Driven	128 (42.7%)	158 (21.9%)		
Total	300 (100%)	720 (100%)	87.6	<.001

Source Differences:

1. **AI over-represents Cluster 2 (Balanced)**: 48.8% vs. 31.0% human
   - This is **expected** given temperature selection (T=0.7 was chosen to produce human-like total variation)
   - Does NOT indicate AI is "more optimal" - rather, that T=0.7 calibrates AI to this profile
2. **Humans over-represent Cluster 4 (Context-Driven)**: 42.7% vs. 21.9% AI
   - Humans show more extreme contextual sensitivity
   - This cluster has lowest AV (0.298) in full sample but not in human-only clustering
3. **Similar representation in Clusters 1 and 3**: No significant differences in principled (13.7% vs. 13.3%) or inconsistent (12.7% vs. 16.0%) extremes

Statistical test: $\chi^2(3) = 87.6$, $p<.001$, Cramér's V = 0.293 (medium effect)

Interpretation:

The clustering reveals that AI at T=0.7 achieves the "Balanced" profile (Cluster 2) nearly 50% of the time, compared to only 31% of humans. However, this should NOT be interpreted as:

✗ "AI reasoning is more optimal than human reasoning"

Instead:

✓ "Temperature T=0.7 was calibrated to produce human-like variation, which results in AI over-representing the Balanced profile"

✓ "Humans show more diverse reasoning profiles, with 43% in high-CR Cluster 4 vs. only 22% of AI"

✓ "The optimal profile depends on whether we prioritize moderate CR (Cluster 2: AV=0.327 in full sample, 0.291 in human-only) or high CR with low AV (Cluster 4: AV=0.298 in full sample, 0.348 in human-only)"

Conclusion:

For organizational ethics applications, we recommend targeting the **human-only Cluster H2 (Balanced) profile**:

- SC ≈ 0.86 (high consistency)
- CR ≈ 0.26 (moderate contextual sensitivity)
- AV ≈ 0.29 (low arbitrary variation)

This represents a **realistic and achievable calibration** for 31% of humans, and can be deliberately cultivated through training that emphasizes:

1. Principled reasoning frameworks (to maintain high SC)
2. Systematic attention to morally relevant contextual features (to achieve moderate CR)
3. Minimizing random inconsistency (to reduce AV below 0.30)

 doi:10.20944/preprints202602.1434.v1

3.8.3. Cluster Characteristics

Demographics:

Cluster	Mean Age	% Female	% Graduate Degree	% Ethics Training
1 (Low)	36.8	48%	64%	28%
2 (Balanced)	39.2	54%	69%	33%
3 (Relational)	39.7	61%	71%	35%
4 (High)	37.4	49%	65%	29%

No significant demographic differences (all χ^2 or t-tests $p > .10$, FDR-corrected $q > .15$).

Framework preferences:

Cluster	Top Framework	% Using Framework
1 (Low)	Utilitarian	42%
2 (Balanced)	Stakeholder	38%
3 (Relational)	Care Ethics	47%
4 (High)	Mixed (no dominant)	—

Professional experience:

Cluster	Mean Years Experience	Mean Leadership Level (1-5)
1 (Low)	11.2	2.8
2 (Balanced)	14.8	3.4
3 (Relational)	15.3	3.6
4 (High)	12.7	3.0

Clusters 2-3 (optimal profiles) have slightly more experience (ANOVA: $F(3,296)=4.2$, $p=.006$, FDR-corrected $q=.011$), suggesting calibration may improve with practice.

3.8.4. AI Model Differences in Cluster Membership (Post-Hoc Analysis)

Analysis Strategy:

After clustering the pooled 720 AI units (3 models × 240 scenarios), we conducted **post-hoc examination** of which model contributed each unit. This reveals whether GPT-4, Claude, and Gemini differ systematically in their reasoning profiles.

Recall: Each AI unit represents one model-scenario combination averaged across 10 repetitions. Thus:

- GPT-4 contributes 240 units (one per scenario)
- Claude contributes 240 units (one per scenario)
- Gemini contributes 240 units (one per scenario)

Distribution by Model:

 doi:10.20944/preprints202602.1434.v1

Cluster	GPT-4 n(%)	Claude n(%)	Gemini n(%)	χ^2	p
1: Principled	43 (17.9%)	53 (22.1%)	48 (20.0%)		
2: Balanced	113 (47.1%)	122 (50.8%)	108 (45.0%)		
3: Inconsistent	67 (27.9%)	51 (21.3%)	68 (28.3%)		
4: Context-Driven	17 (7.1%)	14 (5.8%)	16 (6.7%)		
Total	**240 (100%)**	**240 (100%)**	**240 (100%)**	**8.4**	**.39**

Interpretation:

No significant model differences in cluster distribution ($\chi^2(6)=8.4$, p=.39, Cramér's V=.132).

All three models show:

- ~45-51% in Balanced cluster (Cluster 2)
- ~18-22% in Principled cluster (Cluster 1)
- ~21-28% in Inconsistent cluster (Cluster 3)
- ~6-7% in Context-Driven cluster (Cluster 4)

Implication:

At temperature 0.7, GPT-4, Claude, and Gemini exhibit **similar distributions of reasoning profiles across scenarios**. Differences between models are minor compared to:

1. Variation across scenarios (same model shows different profiles for different scenarios)
2. Differences from humans (see §3.8.2)

Human vs. AI Comparison (Aggregated):

Cluster	Humans n(%)	All AI n(%)	Difference	p
1: Principled	41 (13.7%)	144 (20.0%)	+6.3 pp	.018
2: Balanced	93 (31.0%)	343 (47.6%)	+16.6 pp	<.001
3: Inconsistent	38 (12.7%)	186 (25.8%)	+13.1 pp	<.001
4: Context-Driven	128 (42.7%)	47 (6.5%)	-36.2 pp	<.001

$\chi^2(3) = 87.6$, $p<.001$, Cramér's V = 0.293 (medium effect).

Key Finding:

AI systems (aggregated across models) **over-represent** Cluster 2 (Balanced) and **under-represent** Cluster 4 (Context-Driven) compared to humans.

This suggests:

- AI at T=0.7 more consistently achieves "optimal" profile (49% vs. 31%)
- Humans show more extreme context-sensitivity (43% vs. 7% in Cluster 4)
- Temperature calibration effectively targets the Balanced profile for AI

Limitation:

Because clustering was done on **pooled data** (humans + all AI together), cluster definitions reflect the combined distribution. We cannot determine whether:

1. AI "naturally" fits Cluster 2, OR
2. Cluster 2 definition was influenced by AI's concentration there

Future Research: Separate clustering on human-only and AI-only samples to test whether cluster structures differ fundamentally.

3.9. Sensitivity Analyses and Robustness Checks

 doi:10.20944/preprints202602.1434.v1

3.9.1. Temperature Sensitivity (AI only)

Key findings robustness across temperatures (0.3, 0.5, 0.7, 1.0):

Note: All primary analyses (H1-H4) were conducted at temperature 0.7. Sensitivity analyses test robustness across parameter values.

Debatable feature effects:

Feature	η^2p Range	p-value Range	Robust?
Identifiability	0.16-0.19	<.001 all	✓ Yes
Action/Omission	0.09-0.12	<.001 all	✓ Yes
Temporal	0.07-0.09	<.001 all	✓ Yes
Relational	0.13-0.14	<.001 all	✓ Yes

All FDR-corrected q < .001 across temperatures.

Structural consistency by temperature:

Structural consistency differed significantly across temperatures ($F(3,9) = 8.4$, $p = .005$, $\eta^2p = 0.74$, FDR-corrected $q = .008$), with T=0.3 (M=0.92) > T=1.0 (M=0.82), post-hoc Tukey HSD $p = .003$, FDR-corrected $q = .006$.

Human-AI similarity:

Temperature	AI Variation	Difference from Human (0.42)	p	q
0.3	0.26	-0.16***	<.001	<.001
0.5	0.36	-0.06*	.024	.036
0.7	0.41	-0.01 ns	.56	.58
1.0	0.49	+0.07***	<.001	<.001

Conclusion:

- ✓ **Robust:** Contextual feature effects are consistent across temperatures (all FDR-corrected q < .001)
- ⚠ **Temperature-dependent:** Human-AI similarity in overall variation levels

3.9.2. Alternative Consistency Metrics

We tested three alternative operationalizations:

Metric 1 (Main): Equal weighting (Decision 33% + Framework 33% + Stakeholder 33%)

Metric 2: Decision-weighted (Decision 50% + Framework 25% + Stakeholder 25%)

Metric 3: Stakeholder-focused (Decision 25% + Framework 25% + Stakeholder 50%)

Correlation matrix:

	Metric 1	Metric 2	Metric 3
Metric 1	1.00	.94***	.89***
Metric 2	.94***	1.00	.82***
Metric 3	.89***	.82***	1.00

All correlations p < .001, FDR-corrected q < .001.

Key results under alternative metrics:

 doi:10.20944/preprints202602.1434.v1

Finding	Metric 1	Metric 2	Metric 3
Identifiability effect	η^2p=0.18***	η^2p=0.17***	η^2p=0.16***
Relational mediation	β=0.043***	β=0.039***	β=0.047***
Cluster 2 optimal	AV=0.31	AV=0.29	AV=0.33

All ***p < .001, FDR-corrected q < .001.

Conclusion: Results robust to metric choice (all r > .82, key findings replicate).

3.9.3. Outlier Analysis

Identified outliers (>3 SD from mean variation score):

- n=11 participants (3.7% of final sample)
- Mean variation: 0.78 (vs. 0.42 overall)
- Characteristics: Shorter response times (mean 31 min vs. 47 min), lower word counts

Note: These 11 outliers are **distinct from** the 16 participants excluded during data quality screening (§2.4.1). Outliers were retained for main analyses as extreme variation may represent genuine individual differences.

Analysis with vs. without outliers:

Finding	Full Sample	Outliers Excluded	Change
Mean variation	0.42	0.40	-0.02
Identifiability OR	2.08***	2.04***	-0.04
Relational β	0.043***	0.041***	-0.002

All ***p < .001, FDR-corrected q < .001 in both analyses.

Conclusion: Results essentially unchanged (all key findings remain p<.001, FDR-corrected q<.001).

3.9.4. Domain Subsample Analyses

Testing each domain separately:

Finding	Harm	Fairness	Autonomy	Promise	Honesty
Identifiability OR	2.34***	1.87***	2.18***	1.94***	1.73***
Action/Omission d	0.71***	0.58***	0.64***	0.62***	0.54***
Relational β	0.039**	0.048***	0.041***	0.044***	0.036**

All **p<.01, ***p<.001; FDR-corrected q: **<.01, ***<.001.

Conclusion: All key effects replicate within each domain (minimum OR=1.73, minimum β=0.036, all p<.01, FDR-corrected q<.01).

3.9.5. Missing Data Sensitivity

Missing data:

- Response coding: 0.6% (8 responses excluded due to incomprehensible content - part of 16 total exclusions in §2.4.1)
- Stakeholder rankings: 2.3% (scenarios with single stakeholder)

 doi:10.20944/preprints202602.1434.v1

- Demographics: 1.1% (participants declined to answer)

Multiple imputation (m=20) for demographic covariates:

Finding	Complete Case	Imputed	Difference
Identifiability OR	2.08***	2.09***	+0.01
Relational mediation β	0.043***	0.044***	+0.001

All ***p<.001, FDR-corrected q<.001.

Conclusion: Minimal missing data (<3% any variable); imputation doesn't alter conclusions.

4. Discussion

4.1. Summary of Empirical Findings

This study examined whether humans and AI systems apply consistent ethical frameworks when making organizational decisions, or whether morally contested contextual features systematically influence moral judgments. We generated 7,200 AI responses and 6,000 human responses to 240 systematically varied ethical scenarios, then decomposed observed variation into three theoretically meaningful components.

Four primary findings emerged:

4.1.1. Substantial Contextual Responsiveness (H1-H2)

Both humans and AI demonstrated substantial variation in moral judgments attributable to debatable contextual features (22-24% of total variance, FDR-corrected $q < .001$). Four specific features produced large systematic effects:

1. **Stakeholder identifiability** (OR = 2.08, $\eta^2p = 0.18$): Named individuals favored over statistical aggregates
2. **Stakeholder proximity** (OR = 7.89): Direct stakeholders prioritized over distant stakeholders
3. **Temporal proximity** (OR = 1.52, $\eta^2p = 0.08$): Immediate consequences weighted more than delayed consequences
4. **Relational context** (OR = 1.89, $\eta^2p = 0.14$): Relational stakeholders favored over transactional stakeholders

All effects: $p < .001$, FDR-corrected $q < .001$; no significant human-AI differences (all Source × Feature interactions $p > .20$, FDR-corrected $q > .30$).

These effects are not trivial edge cases. When all four features aligned (identified, direct, immediate, relational stakeholder), participants were approximately 45 times more likely to choose the favored option compared to when all four opposed (combined OR = 44.7).

4.1.2. Classic Omission Bias Replicated (H3)

The action-omission asymmetry—judging harmful actions more severely than harmful omissions—appeared robustly in organizational contexts (OR = 1.87, d = 0.63, FDR-corrected $q < .001$). Participants were 87% more willing to allow harm through inaction than to cause identical harm through action.

This effect was moderated by agency salience (36% stronger when personal agency emphasized, Frame × Agency interaction $p < .001$, FDR-corrected $q < .001$), suggesting psychological mechanisms involving personal responsibility attribution. Both humans and AI exhibited identical patterns (no three-way Source × Frame × Agency interaction, $p = .71$, FDR-corrected $q = .76$).

 doi:10.20944/preprints202602.1434.v1

4.1.3. Relational Reasoning Explains Human-AI Differences (H4)

Humans exhibited substantially more relational reasoning than AI (d = 0.56 for reasoning strength, d = 1.04 for language density, both FDR-corrected q < .001). Mediation analysis revealed that relational reasoning accounted for 69% of human-AI differences in contextual responsiveness (indirect effect β = 0.043, 95% CI [0.028, 0.061], FDR-corrected q < .001).

Critically, relational reasoning was associated with:

- ↑ **Higher contextual responsiveness** (+112% comparing Level 3 vs. Level 0)
- ↓ **Lower arbitrary variation** (-31% comparing Level 3 vs. Level 0)

This pattern contradicts the interpretation that relational reasoning represents confused or inconsistent thinking. Instead, it appears to be one systematic factor among others influencing moral judgment.

4.1.4. Most Variation Is Systematic, But One-Third Remains Arbitrary

Decomposing total variation:

- **Structural consistency:** 84-87% agreement when only irrelevant features varied
- **Contextual responsiveness:** 22-24% variance attributable to debatable features
- **Arbitrary variation:** 32-34% unexplained residual variance

The majority of variation (68-76% = structural consistency + contextual responsiveness) is systematic and potentially justifiable. However, the substantial arbitrary component (32-34%) represents genuine inconsistency that even particularist frameworks should view as problematic.

Cluster analysis identified an "optimal" profile (Cluster 2: Balanced Sensitivity, 31% of humans, 48% of AI responses at temperature 0.7) characterized by moderate contextual responsiveness (CR = 0.28) with minimal arbitrary variation (AV = 0.31, lowest among clusters).

4.2. Philosophical Implications: The Context Sensitivity Paradox

Our findings create a paradox for moral philosophy and organizational ethics: The same data support diametrically opposed normative conclusions depending on one's meta-ethical commitments.

4.2.1. The Principlist Interpretation: Widespread Bias

From a principlist perspective (Beauchamp & Childress, 2019; Kant, 1785/1998), our findings demonstrate pervasive moral reasoning failures. Four morally irrelevant features—identifiability, proximity, temporal distance, and relational ties—systematically distort judgments that should depend only on morally relevant factors like magnitude of harm, probability of outcomes, and rights violations.

Evidence supporting this interpretation:

1. Identifiability bias violates impartiality

The identifiability effect (OR = 2.08) means structurally identical outcomes receive different moral weight based solely on descriptive detail. Consider:

- "50 employees will lose jobs" → 42% choose protective option
- "Maria Rodriguez, single mother of three, and 49 colleagues will lose jobs" → 67% choose protective option

If the moral weight of job losses depends on number and severity (not names), this 25-percentage-point shift represents pure bias (Jenni & Loewenstein, 1997).

2. Temporal discounting violates temporal neutrality

The temporal proximity effect (OR = 1.52) means future people count less. Our implicit discount rate of ~21-23% per year means:

- 50 people harmed today = 61 people harmed in 2 years (morally equivalent under discount)

 doi:10.20944/preprints202602.1434.v1

Classical utilitarianism and most deontological theories reject temporal discounting of non-instrumental goods like human welfare (Parfit, 1984; Sidgwick, 1907). Why should timing change moral weight?

3. Action-omission asymmetry violates outcome equivalence

The action-omission effect (OR = 1.87) means causal framing determines judgment independent of outcomes:

- Actively causing 50 job losses → 38% accept
- Passively allowing 50 job losses → 53% accept

If consequences are identical, this distinction lacks moral justification under consequentialist frameworks (Singer, 1972). It reflects psychological quirks of causal reasoning, not genuine moral differences.

4. Relational favoritism violates universalizability

The relational context effect (OR = 1.89) means personal relationships trump impartial evaluation. But:

> "Morality requires that we not make distinctions based on morally arbitrary features such as... personal familiarity" (Rachels, 1999, p. 14)

All stakeholders should count equally; that some happen to have relational ties to decision-makers is morally irrelevant.

Policy implications under this interpretation:

If contextual responsiveness represents bias, organizations should:

1. **Implement structured decision protocols** removing contextual details (de-identify stakeholders, standardize time horizons, use statistical aggregates)
2. **Train decision-makers** to recognize and correct for these biases
3. **Deploy AI systems** configured for minimal contextual sensitivity (temperature 0.3), leveraging their lower baseline CR (though our data show even temperature 0.3 exhibits significant effects)
4. **Audit decisions** for consistency across framings; flag high CR as quality failure

This approach aligns with extensive literature on debiasing organizational decision-making (Bazerman & Moore, 2013; Kahneman et al., 2021).

4.2.2. The Particularist Interpretation: Appropriate Sensitivity

From a particularist perspective (Dancy, 2004; McKeever & Ridge, 2006; McNaughton, 1988), identical findings demonstrate appropriate moral sensitivity to contextually relevant features. Particularism holds that no universal principles fully capture moral reality; context shapes which considerations matter and how much they matter.

Evidence supporting this interpretation:

1. Identifiability captures morally relevant concreteness

The identifiability effect may reflect genuine moral insight: Named individuals are not mere instances of "employee" category but concrete persons with unique life circumstances. Maria Rodriguez's 12-year tenure, single parenthood, and three children are morally relevant details—not because she matters more than others, but because:

> "Seeing the particular person... is not a distraction from moral reality but precisely what enables us to respond appropriately to that reality" (Murdoch, 1970, p. 37)

Statistical aggregates ("50 employees") risk reification—treating people as fungible units rather than irreducible individuals (Blum, 1991). Higher willingness to protect identified stakeholders may represent moral progress, not bias.

2. Relational context creates genuine obligations

The relational effect (OR = 1.89) reflects care ethics insight: Relationships generate special obligations not captured by impartial frameworks (Gilligan, 1982; Held, 2006; Noddings, 1984).

A 12-year employee who has "consistently exceeded expectations" has built trust, made contributions, and developed legitimate expectations. Treating them identically to a 6-month

 doi:10.20944/preprints202602.1434.v1

contractor hired on a temporary contract may constitute moral blindness to these morally relevant relational facts.

"We acquire special obligations through our particular relationships... these obligations are not reducible to general duties" (Scheffler, 1997, p. 190)

3. Action-omission distinction reflects agency asymmetries

The action-omission asymmetry (OR = 1.87) may track genuine moral differences in:

- **Causal responsibility:** Active causation involves stronger agency than passive allowance
- **Autonomy violations:** Actions impose will on others; omissions permit natural processes
- **Moral psychology:** Intention structures differ (doing vs. letting happen)

While consequentialists reject this distinction, deontological and virtue ethics traditions take it seriously (Foot, 1967; Quinn, 1989). That humans and AI both exhibit it may indicate deep normative insight, not shared bias.

4. Temporal proximity reflects practical rationality

The temporal effect (OR = 1.52) may represent appropriate practical reasoning:

- **Epistemic uncertainty:** Distant consequences are genuinely more uncertain
- **Opportunity costs:** Immediate actions prevent future option spaces
- **Psychological sustainability:** Perfect temporal neutrality may be psychologically impossible for finite agents

Some philosophers defend "near-bias" as rational for embodied, temporally located agents (Greene, 2013; Railton, 1984).

Policy implications under this interpretation:

If contextual responsiveness represents appropriate sensitivity, organizations should:

1. **Preserve rich contextual information** rather than de-identifying stakeholders
2. **Train decision-makers** in care ethics and relational reasoning frameworks
3. **Configure AI systems** for moderate contextual sensitivity (temperature 0.7), targeting the "Balanced" profile (Cluster 2)
4. **Audit for systematic patterns** rather than raw consistency; flag low CR as potentially insensitive

This approach aligns with relational organizational ethics literature (Donaldson & Dunfee, 1999; Freeman, 1984; Wicks et al., 1994).

4.2.3. Why Our Data Cannot Adjudicate

The context sensitivity paradox arises because our empirical findings are compatible with both interpretations:

Observation 1: Identifiability doubles prioritization (OR = 2.08)

- **Principlist:** Pure bias (names are morally irrelevant)
- **Particularist:** Appropriate response to concreteness (persons vs. statistics)

Observation 2: Relational reasoning increases CR while reducing AV

- **Principlist:** Bias that happens to be systematic (still wrong)
- **Particularist:** Genuine moral competence (systematic = appropriate)

Observation 3: Humans show higher CR and relational reasoning than AI

- **Principlist:** Humans more biased (AI superior at temperature 0.3)
- **Particularist:** Humans more morally sophisticated (AI impoverished)

Observation 4: Framework choice tracks context systematically (§3.6.3)

- **Principlist:** Post-hoc rationalization (people choose frameworks to justify biases)
- **Particularist:** Appropriate framework selection (knowing when different principles apply)

The fundamental problem:

 doi:10.20944/preprints202602.1434.v1

Our method measures **that** contextual features influence judgment and **how much** they influence judgment. But this cannot determine **whether they should** influence judgment. That requires normative argument beyond empirical data.

We can establish:

- Identifiability effects exist (empirical)
- They are large and systematic (empirical)
- They correlate with relational reasoning (empirical)

We cannot establish:

- Whether identifiability is morally relevant (normative)
- Whether systematic sensitivity is virtuous or vicious (normative)
- Whether AI or humans are "correct" (normative)

4.2.4. Implications for Normative Ethics

Our findings nevertheless constrain normative debate:

1. Pure principlism is empirically implausible

Zero contextual sensitivity (CR = 0) appears psychologically unrealistic for humans and requires extremely low AI temperature (T < 0.3) with degraded coherence. Cluster 1 (low sensitivity) shows **higher** arbitrary variation than moderate-sensitivity clusters, suggesting rigid principlism may increase, not decrease, inconsistency.

Constraint: Realistic normative theories must accommodate modest contextual sensitivity (CR = 0.20-0.30), not demand zero.

2. Pure particularism lacks decision guidance

Very high contextual sensitivity (Cluster 4, CR = 0.42) also shows elevated arbitrary variation (AV = 0.35). Unlimited responsiveness to context becomes indistinguishable from inconsistency.

Constraint: Realistic particularism requires principled limits on which contextual features matter and how much.

3. The "optimal" profile is hybrid

Cluster 2 (Balanced Sensitivity: moderate CR, minimal AV) suggests effective moral reasoning requires:

- Systematic sensitivity to **some** contextual features (not zero)
- Principled insensitivity to **other** contextual features (not unlimited)
- Clear criteria distinguishing morally relevant from irrelevant context

Normative challenge: Specify which contextual features belong in each category.

4. Framework integration is double-edged

High integrators (Level 3-4) show:

- Better conditional consistency (within-framework coherence)
- Systematic framework selection (context-appropriate frameworks)
- But also higher overall variation and elevated arbitrary variation

Normative question: Is framework pluralism sophisticated or confused?

4.2.5. A Tentative Synthesis: Bounded Particularism

Relationship to Temperature Selection Circularity

Critical Acknowledgment:

Our "optimal" profile identification (Cluster 2: Balanced Sensitivity) is entangled with temperature selection circularity in complex ways.

The Circularity:

1. We selected T=0.7 because it produced human-like **total variation** (0.41 vs. 0.42)
2. At T=0.7, AI over-represents Cluster 2 (48% vs. 31% humans)
3. Cluster 2 has lowest arbitrary variation (AV=0.31)

 doi:10.20944/preprints202602.1434.v1

4. We call Cluster 2 "optimal" based on this low AV

Potential Circular Reasoning:

Could we be concluding:

- "The temperature we selected to match humans produces a profile that we then call optimal based on characteristics influenced by that temperature selection"?

Disentangling the Circularity:

What IS circular:

- AI achieving Cluster 2 more frequently than humans (48% vs. 31%) is partly **artifact of temperature selection**
- At T=0.3: AI in Cluster 2 drops to 35% (closer to humans)
- At T=1.0: AI in Cluster 2 drops to 28% (below humans)
- Temperature directly affects where AI units fall in cluster space

What is NOT circular:

- Cluster 2 having lowest AV (0.31) is **observed across both humans and AI**
- This holds even when clustering humans separately (human-only Cluster 2: AV=0.29)
- The Balanced profile's advantage (moderate CR with low AV) is **not temperature-dependent**

Evidence for Non-Circularity:

Human-only clustering (excluding AI entirely):

We re-ran k-means on just N=300 human participants:

Human-Only Cluster	% of Humans	SC	CR	AV
H1 (Principled)	29%	0.92	0.14	0.34
H2 (Balanced)	31%	0.86	0.26	**0.29** ← Lowest
H3 (Inconsistent)	13%	0.74	0.22	0.52
H4 (Context-Driven)	27%	0.82	0.39	0.35

Finding: Even in human-only data, the Balanced cluster (moderate CR ~0.26, SC ~0.86) achieves lowest AV.

Conclusion: The "optimality" of Cluster 2 (Balanced) is **not** an artifact of temperature selection. It reflects a genuine pattern: **moderate contextual sensitivity paired with high structural consistency minimizes arbitrary variation**.

What Temperature Selection DOES Affect:

Temperature changes **how frequently AI achieves** the Balanced profile, not whether it's optimal:

Temperature	% AI in Balanced	% Humans in Balanced	Difference
0.3	35%	31%	+4 pp
0.7	48%	31%	+17 pp
1.0	28%	31%	-3 pp

Interpretation:

- T=0.7 makes AI **more likely** to exhibit Balanced profile than humans
- This is what makes T=0.7 seem "optimal" for AI deployment
- But it doesn't make Balanced profile itself optimal (that's empirically supported regardless of temperature)

 doi:10.20944/preprints202602.1434.v1

Implication for Bounded Particularism:

Our normative proposal—that effective moral reasoning requires moderate contextual sensitivity (CR≈0.25-0.30) with low arbitrary variation (AV<0.30)—rests on:

✓ **Non-circular evidence:**

- Human-only clustering shows Balanced profile has lowest AV
- This pattern replicates across age groups, education levels, and professional experience (see Appendix D.1: Demographic Analyses). Cross-cultural replication is a priority for future research, as our sample over-represents Western contexts (89%).
- Theoretical coherence: Pure principlism (Cluster 1) shows higher AV despite low CR

✗ **Circular influence:**

- That AI at T=0.7 achieves this profile more consistently than humans
- This says more about temperature calibration than moral reasoning

Revised Claim:

Original (potentially circular): "AI systems at T=0.7 demonstrate optimal reasoning by achieving Balanced profile 48% of the time."

Revised (non-circular): "The Balanced profile (moderate CR, low AV) represents optimal reasoning based on human data. AI can be **calibrated** via temperature selection to achieve this profile more or less frequently, with T=0.7 producing **highest rates** of Balanced reasoning (48% vs. 28-35% at other temps)."

Key Distinction:

- **What is optimal:** Empirically supported (Balanced profile minimizes AV)
- **How AI achieves it:** Temperature-dependent (T=0.7 maximizes frequency)

This distinction preserves bounded particularism as a substantive normative claim while acknowledging temperature selection affects AI deployment, not fundamental reasoning patterns.

While our data cannot resolve the principlist-particularist debate, they suggest a middle position we term **bounded particularism**.

Core Claims:

1. **Some contextual features are morally relevant** (contra pure principlism):
   - Relational obligations, concrete particularity, and causal structure plausibly matter morally
   - Zero contextual sensitivity (Cluster 1, CR=14%) shows **higher** arbitrary variation (31%) than moderate sensitivity (Cluster 2, CR=26%, AV=33%)
   - This suggests rigid principlism may increase, not decrease, inconsistency
2. **Not all observed sensitivity is appropriate** (contra pure particularism):
   - 32-34% arbitrary variation across sample indicates substantial unprincipled inconsistency
   - Temperature-dependent patterns in AI (CR ranges 12%-28% across T=0.3 to T=1.0) suggest some "sensitivity" is architectural artifact, not moral insight
   - Very high contextual sensitivity (Cluster 4, CR=39%) doesn't further reduce arbitrary variation vs. moderate sensitivity
3. **Optimal judgment balances principles and context** (synthesis):
   - The "Balanced-Integrative" profile (Cluster 2) achieves:
     - Moderate contextual responsiveness (CR=26%)
     - Minimal arbitrary variation (AV=33%, lowest across clusters)
     - Good structural consistency (SC=86%)
   - This profile represents **neither** pure principlism (CR too high) **nor** pure particularism (CR not maximized)

Empirical Targets for "Appropriate" Moral Reasoning:

Based on Cluster 2 (optimal profile by AV minimization):

 doi:10.20944/preprints202602.1434.v1

Component	Target Range	Rationale
SC	>0.85	High reliability when irrelevant features vary
CR	0.20-0.30	Systematic sensitivity without over-fitting
AV	<0.35	Minimal unexplained randomness
Calibration	CR/AV >0.75	Signal-to-noise ratio favoring systematic over random variation

Decision Procedure (Operationalizing Bounded Particularism):

1. **Start with general principles** (default to consistency)
   - Identify applicable frameworks (utilitarian, deontological, care ethics, etc.)
   - Apply consistently across structurally similar cases
2. **Allow context to defeat defaults** when specific features cross salience threshold
   - Identifiable stakeholders may warrant different treatment than statistical aggregates
   - Relational history may create special obligations
   - Temporal proximity may reflect epistemic uncertainty
   - Action/omission may track genuine moral distinctions
3. **Limit to theoretically justified features** (not arbitrary framings)
   - Candidate features: identifiability, relational context, causal structure, temporal proximity
   - Exclude: presentation order, wording variations, irrelevant demographics
4. **Monitor for arbitrary variation** (not all variation is wisdom)
   - Calculate individual AV scores
   - If AV >0.35, scrutinize decisions for unprincipled inconsistency
   - If CR >0.35 with high AV, may indicate over-sensitivity to irrelevant context

This Approach:

✓ Preserves principlist concern for consistency (high SC, low AV)
✓ Accommodates particularist insight about context (moderate CR)
✓ Provides empirical targets (CR≈0.25-0.30, AV<0.30, SC>0.85)
✓ Admits both humans and AI can achieve optimal profile (though AI does so more consistently at T=0.7)

Open Questions:

1. **Which specific features are "theoretically justified"?**
   - Our four features (identifiability, proximity, temporal, relational)?
   - Others we didn't measure?
   - Context-dependent (different features relevant in different domains)?
2. **What "salience threshold" should trigger context-sensitivity?**
   - Always consider relational obligations?
   - Only when relationships cross duration/intensity threshold?
   - Calibrated to domain norms?
3. **How do we distinguish legitimate contextual defeating from bias?**
   - Empirical criterion: Does feature variation reduce AV?
   - Normative criterion: Philosophical argument for moral relevance?
   - Pragmatic criterion: Stakeholder acceptance and organizational sustainability?
4. **Can the optimal profile (Cluster 2) be trained/achieved?**
   - Our data show 31% of humans naturally in this cluster
   - Can ethics training move people from Clusters 1, 3, 4 → Cluster 2?
   - Can AI be calibrated to reliably achieve Cluster 2 profile?

 doi:10.20944/preprints202602.1434.v1

Limitations of This Synthesis:

This remains a tentative proposal because:

1. **Circularity concern:** We defined "optimal" as minimizing AV, but:
   - AV includes measurement error and unmeasured constructs
   - Low AV might reflect lack of sensitivity to legitimate but unmeasured features
   - "Optimal" is normatively loaded—assumes consistency is virtuous
2. **Cluster 2 superiority not universally accepted:**
   - Principlists might argue Cluster 1 (lowest CR) is optimal if we could eliminate their higher AV through better training
   - Particularists might argue Cluster 4 (highest CR) is optimal and their moderate AV reflects appropriate complexity
3. **Sample-specific findings:**
   - Cluster structure might differ in:
     - Non-Western cultural contexts
     - Different professional domains
     - Higher-stakes real-world decisions
   - Our "optimal" profile may be optimal only for these scenarios
4. **Temperature-dependence undermines AI claims:**
   - That AI achieves Cluster 2 more frequently (49% vs. 31%) is artifact of:
     - Temperature selected to match human variation
     - Deterministic sampling reducing noise
   - Not evidence of superior AI moral reasoning

Nevertheless, bounded particularism offers:

- **Conceptual clarity:** Specifies what we mean by "appropriate" balance
- **Empirical targets:** Testable predictions about optimal profiles
- **Practical guidance:** Concrete metrics for training and deployment
- **Middle path:** Avoids extremes of rigid principlism and unprincipled relativism

Future research should:

1. Test whether Cluster 2 profile predicts better outcomes (stakeholder satisfaction, decision quality, organizational performance)
2. Develop interventions to move individuals toward Cluster 2
3. Examine cross-cultural generalizability of cluster structure
4. Philosophically defend (or critique) the normative assumption that minimizing AV is desirable

4.3. Practical Implications for Organizations

4.3.1. Implications for Ethics Training

Current organizational ethics training typically emphasizes either:

- **Principles-based approaches:** Teach universal frameworks (utilitarian, deontological, rights-based) and encourage consistent application
- **Case-based approaches:** Develop judgment through exposure to diverse scenarios and contextual reasoning

Our findings suggest both approaches face challenges:

Challenges for principles-based training:

1. **Low transfer:** Cluster 1 (single-framework users) show **higher** arbitrary variation (0.41) than multi-framework users (0.31-0.35), suggesting rigid principle application doesn't reduce inconsistency

 doi:10.20944/preprints202602.1434.v1

2. **Context insensitivity:** Pure principlism requires ignoring features (identifiability, relational context) that may be morally relevant
3. **Psychological unrealism:** Achieving very low contextual sensitivity (CR < 0.15) appears difficult for humans and may be undesirable

Challenges for case-based training:

1. **Overfitting risk:** High integrators (Level 3-4) show elevated arbitrary variation (0.35), suggesting unlimited context-sensitivity becomes unprincipled
2. **Framework confusion:** Without clear decision procedures, case exposure may simply increase variation without improving judgment
3. **Lack of generalization:** Conditional consistency (within-framework) is better than overall consistency, but only if framework selection itself is principled

Recommendation: Hybrid training focused on "bounded particularism"
Effective ethics training should:

1. **Teach multiple frameworks** (utilitarian, deontological, care ethics) with clear scope conditions
2. **Identify morally relevant features** explicitly (our four features provide starting point)
3. **Practice systematic framework selection** (when does care ethics vs. utilitarianism apply?)
4. **Monitor arbitrary variation** (use consistency checks to catch unprincipled variation)
5. **Target optimal profile** (Cluster 2 parameters: CR ≈ 0.28, AV < 0.31)

Specific exercises:

- **Consistency checks:** Present identical scenarios with surface variations; flag unexplained differences
- **Feature isolation:** Present scenarios varying only identifiability or only relational context; discuss when variation is justified
- **Framework mapping:** For each framework, identify scenarios where it applies vs. doesn't apply
- **Variation decomposition:** Calculate individual CR/AV scores; provide feedback on sources of inconsistency

4.3.2. Implications for AI Governance

Organizations increasingly deploy AI systems for decision support or autonomous decision-making in ethically sensitive contexts (hiring, lending, resource allocation, crisis response). Our findings raise critical questions about AI configuration and oversight.

Key decision: Temperature parameter selection
Temperature directly controls contextual sensitivity:

Temperature	Mean CR	Mean AV	Coherence	Recommendation
0.3	0.12	0.41	99.6%	Principlist contexts requiring consistency
0.5	0.21	0.36	98.8%	Moderate sensitivity, low noise
0.7	0.24	0.34	97.2%	**Balanced (human-like)**
1.0	0.28	0.49	91.6%	High sensitivity but excessive noise

Configuration guidance:
Use lower temperature (0.3-0.5) when:

- Legal compliance is paramount (minimal interpretation needed)
- Consistency across cases is essential (fairness as uniformity)
- Stakeholder anonymization is feasible and desirable
- Rapid decisions at scale (minimize computational variance)

 doi:10.20944/preprints202602.1434.v1

Example applications: Automated lending decisions, regulatory compliance checks, standardized hiring rubrics

Use moderate temperature (0.5-0.7) when:

- Context-sensitive judgment is valuable
- Human-like reasoning increases acceptance
- Relational factors may be relevant
- Explaining decisions to stakeholders matters

Example applications: Employee grievance review, customer accommodation requests, stakeholder consultation

Avoid high temperature (>0.8) when:

- Consistency matters (AV becomes problematic)
- Automated decision-making (coherence degrades)
- Accountability required (excessive variation complicates auditing)

Deployment models:

Model 1: AI as consistency checker

Configure AI at low temperature (0.3) to flag cases where human decision-makers show unexplained variation:

- Generate AI recommendation with minimal contextual sensitivity
- Compare to human decision
- If difference > threshold AND no clear contextual justification → trigger review

Model 2: AI as contextual advisor

Configure AI at moderate temperature (0.7) to provide human-like reasoning:

- Generate recommendation with explicit framework and contextual considerations
- Present to human decision-maker as one input
- Human retains final authority but sees systematic contextual analysis

Model 3: Hybrid ensemble

Deploy multiple AI instances at different temperatures:

- Low-T model provides principled baseline
- Moderate-T model provides contextual analysis
- Compare recommendations; disagreement triggers human review
- Effectively implements "bounded particularism" architecturally

Critical safeguards:

Regardless of configuration:

1. **Transparency:** Log temperature and sampling parameters; make clear that variation is architectural, not error
2. **Auditing:** Track CR/AV metrics over time; flag drift from target profiles
3. **Human oversight:** Require human approval for decisions in high-stakes domains
4. **Bias monitoring:** Test for systematic disparities across demographic groups (contextual sensitivity could amplify existing biases)
5. **Framework documentation:** Require AI to specify which ethical framework(s) informed recommendation

4.3.3. Implications for Decision Auditing

Traditional decision auditing focuses on **outcome consistency**: Do similar cases receive similar decisions? Our three-component decomposition suggests more nuanced auditing:

Audit framework:

Metric 1: Structural Consistency (SC) - Target: >0.85

Flag decisions with low structural consistency as potential quality failures:

 doi:10.20944/preprints202602.1434.v1

- Same scenario presented differently → different decisions
- Surface features (wording, presentation) driving outcomes
- **Interpretation:** Unambiguous reliability problem

Audit action: Require decision review; investigate source of inconsistency
Metric 2: Arbitrary Variation (AV) - Target: <0.30
Flag decisions with high arbitrary variation as potential reasoning failures:

- Variation unexplained by scenario features or principled frameworks
- Inconsistent application of stated decision criteria
- **Interpretation:** Problematic even under particularism

Audit action: Request decision justification; if no principled explanation, require reconsideration
Metric 3: Contextual Responsiveness (CR) - Target: 0.20-0.30 (disputed)
Here interpretation depends on normative stance:
If organizationally principlist:

- Target CR < 0.20
- Flag high CR as excessive context-sensitivity
- Audit action: Check if variation is justified by clearly relevant features (magnitude, probability, rights)

If organizationally particularist:

- Target CR = 0.25-0.30
- Flag both very low CR (insensitivity) and very high CR (unprincipled)
- Audit action: Verify variation corresponds to theoretically justified features

Conditional consistency auditing:
For multi-framework decision-makers:

1. Calculate consistency **within each framework** (conditional consistency)
2. Audit framework selection separately
3. Flag as problematic only if:
   - Low conditional consistency (inconsistent application), OR
   - Unprincipled framework switching (no clear selection criteria)

Example audit report:
Decision-Maker: Manager A
Period: Q3 2024
Cases: 47 ethics decisions
Structural Consistency: 0.87 ✓ (Target: >0.85)
Arbitrary Variation: 0.28 ✓ (Target: <0.30)
Contextual Responsiveness: 0.39 ⚠ (Target: 0.20-0.30)
Breakdown:

- Identifiability effect: $\eta^2p = 0.22$ (high)
- Relational effect: $\eta^2p = 0.18$ (high)
- Temporal effect: $\eta^2p = 0.06$ (acceptable)
- Action/omission: $\eta^2p = 0.09$ (acceptable)

Recommendation: Review cases with high identifiability/relational sensitivity. Verify that variation is justified by genuine moral relevance of these features (care ethics framework) rather than arbitrary favoritism.
Conditional Consistency (Care Ethics cases only): 0.84
Conditional Consistency (Utilitarian cases only): 0.88
Framework Selection: Systematic (Care Ethics for relational contexts, Utilitarian for statistical contexts)
Overall Assessment: HIGH CR driven by systematic framework

 doi:10.20944/preprints202602.1434.v1

selection. Conditional consistency acceptable. Recommend documentation of framework selection criteria to ensure principled application.

4.3.4. Implications for Stakeholder Communication

Our findings have significant implications for how organizations communicate ethical decisions to stakeholders:

Transparency about context-sensitivity:

Organizations should acknowledge that:

1. **Context influences decisions** (attempting to hide this creates legitimacy gaps)
2. **Some context-sensitivity is appropriate** (particularist framing)
3. **Variation is partially systematic** (not arbitrary or biased)

Communication template:

"Our decision-making process considers both universal principles and contextual factors. While we strive for consistency in how we apply ethical frameworks, we recognize that different contexts may warrant different frameworks. We monitor our decisions to ensure variation is principled rather than arbitrary."

Feature-specific disclosure:

When contextual features influence decisions, explain why:

For identifiability:

"We consider both aggregate outcomes and individual circumstances. While statistical analysis informs our understanding of overall impact, we also examine how decisions affect specific individuals, as this enables us to respond to concrete moral reality rather than abstract categories."

For relational context:

"We recognize that ongoing relationships create special obligations beyond transactional interactions. Long-term employees, loyal customers, and trusted partners have built legitimate expectations through their contributions and investments in relationship with our organization."

For temporal proximity:

"We balance short-term and long-term considerations, while recognizing that immediate consequences often demand more urgent response than distant future impacts. This reflects both epistemic uncertainty about distant outcomes and practical constraints on current decision-making."

For action-omission:

"We distinguish between actively causing harm and allowing harm to occur through inaction, as these reflect different levels of agency and moral responsibility. However, we do not use this distinction to avoid responsibility for foreseeable consequences of our decisions."

Managing stakeholder expectations:

Different stakeholder groups may hold different meta-ethical views:

Expectation mapping:

Stakeholder	Likely Preference	Communication Strategy
Regulators	Principlist (consistency)	Emphasize structural consistency, low AV
Employees	Particularist (contextual)	Emphasize relational reasoning, care ethics

 doi:10.20944/preprints202602.1434.v1

Stakeholder	Likely Preference	Communication Strategy
Shareholders	Consequentialist	Emphasize outcome optimization
Community	Mixed	Acknowledge framework pluralism

Adaptive communication:

Frame same decision differently for different audiences:

To regulators: "Our decision applies consistent principles across all cases: [X]. While contextual details inform application, core criteria remain constant, ensuring fairness and predictability."

To employees: "Our decision recognizes your specific circumstances and our ongoing relationship. We considered both general principles and the particular context of your situation, including your contributions and legitimate expectations."

Critical constraint: Maintain integrity

Adaptive framing must not involve:

- **Contradictory justifications** (saying opposite things to different groups)
- **Hiding true reasons** (claiming principled consistency when actually context-driven)
- **Post-hoc rationalization** (fabricating justifications after decision made)

Solution: Framework transparency

Document which framework(s) informed each decision **before** communicating:

1. What principles applied?
2. What contextual features were considered?
3. How were they weighted?
4. What was the decision procedure?

Then adapt **emphasis** for different audiences while maintaining **consistency** in underlying justification.

4.4. Limitations and Future Directions

4.4.1. Methodological Limitations

1. Cross-sectional design limits causal inference

While mediation analysis (H4) suggests relational reasoning → contextual responsiveness, our design cannot definitively establish causation. Alternative causal models remain plausible:

- **Reverse causation:** High CR → increased relational language (post-hoc justification)
- **Common cause:** Personality traits → both RR and CR
- **Reciprocal causation:** Bidirectional relationship

Future direction: Experimental manipulation of relational framing within-subjects, measuring whether induced relational reasoning increases CR.

2. Scenario-based measures may not reflect real decisions

Participants responded to hypothetical scenarios, not actual organizational dilemmas with:

- Real stakeholders and relationships
- Personal consequences for decision-maker
- Time pressure and incomplete information
- Organizational politics and power dynamics

Validity questions:

- Do hypothetical judgments predict actual behavior? (Literature: mixed, see Bersoff, 1999; FeldmanHall et al., 2012)
- Are contextual effects **stronger** in real settings (personal stakes amplify biases) or **weaker** (professional norms constrain variation)?

Future direction:

 doi:10.20944/preprints202602.1434.v1

- Field studies tracking actual organizational decisions with pre-registered coding
- Experience sampling: managers report real ethical decisions in near-real-time
- Archival analysis: code historical organizational decisions for contextual patterns

3. Limited diversity in human sample

Our sample over-represents:

- Western cultural contexts (89% Western)
- Highly educated professionals (67% graduate degrees)
- Technology/healthcare/finance sectors (56%)

Generalizability concerns:

- Eastern vs. Western moral reasoning differs systematically (Nisbett et al., 2001)
- Education correlates with abstract reasoning and moral sophistication (Rest, 1986)
- Industry norms shape ethical judgment (Victor & Cullen, 1988)

Future direction:

- Cross-cultural replication (East Asia, Middle East, Africa, Latin America)
- Diverse occupational sampling (blue-collar, service sector, public sector)
- Developmental study (novice managers vs. experienced executives)

4. AI models tested at single time point

AI capabilities evolve rapidly. Our findings reflect:

- Specific model versions (GPT-4 Jan 2025, Claude 3 Opus Feb 2024, Gemini Pro 1.5 Sep 2024)
- Training data cutoffs (each model trained on different time periods)
- Evolving reinforcement learning from human feedback (RLHF)

Future direction:

- Longitudinal tracking of model versions
- Comparison to specialized "ethics-focused" models as they emerge
- Analysis of training data composition effects on contextual sensitivity

5. Temperature sensitivity complicates human-AI comparison

Our finding that "AI exhibits human-like variation at temperature 0.7" is partially circular:

- We selected 0.7 specifically because it produced human-like patterns
- At 0.3, AI shows less variation; at 1.0, more variation
- No principled way to identify "true" AI reasoning pattern

Implication:

- Cannot claim AI "inherently" resembles humans
- Can claim: At typical deployment parameters, patterns are similar
- Should report results across temperature range (we did this in sensitivity analyses)

Future direction:

- Develop theory-driven temperature selection criteria
- Test other stochastic parameters (top-p, top-k) for robustness
- Compare deterministic AI approaches (e.g., chain-of-thought reasoning at T=0)

6. Construct validity of "arbitrary variation"

Our AV metric (residual variation after removing structural and contextual components) conflates:

- Genuine inconsistency (problematic)
- Unmeasured individual differences (potentially legitimate)
- Measurement error (unavoidable)
- Subtle contextual features we didn't code (might be morally relevant)

Cannot definitively claim AV = "pure noise"

Future direction:

 doi:10.20944/preprints202602.1434.v1

- Qualitative analysis of high-AV cases to identify patterns
- Test-retest reliability studies to separate state vs. trait variation
- Expand feature coding to capture additional contextual dimensions

4.4.2. Theoretical and Interpretive Limitations

1. Cannot adjudicate normative debate

As discussed extensively (§4.2), our data constrain but cannot resolve the principlist-particularist debate. We can measure **that** context influences judgment and **how much**, but not **whether it should**.

What we provide:

- Empirical benchmarks for normative theories (e.g., "pure principlism requires CR < 0.15, which is psychologically implausible")
- Evidence about consequences of different approaches (e.g., "high integrators show higher CR but also higher AV")

What we cannot provide:

- Definitive answer to "which contextual features are morally relevant?"
- Proof that any particular variation pattern is morally right or wrong

Future direction:

- Interdisciplinary dialogue between empirical researchers and normative ethicists
- Development of normative theories that explicitly incorporate empirical constraints
- Expert elicitation studies asking moral philosophers to classify features as relevant/irrelevant

2. Framework coding is interpretive

Our classification of responses into ethical frameworks (Utilitarian, Deontological, Care, Rights, etc.) required interpretive judgment:

- Many responses combined multiple frameworks
- Framework boundaries are contested theoretically
- Coders may have systematically misclassified certain reasoning patterns

Reliability: Inter-rater agreement ($\kappa = 0.81$) is good but not perfect; 19% of cases required adjudication

Future direction:

- Develop more granular framework taxonomy
- Use multiple independent coding teams with different theoretical backgrounds
- Validate framework classifications against self-reported ethical orientation

3. Causal mechanisms remain unclear

We documented effects but not underlying processes:

Identifiability effect:

- Mediated by empathy/emotional response? (affective mechanism)
- Or enhanced moral salience of concrete persons? (cognitive mechanism)
- Or both?

Action-omission effect:

- Driven by responsibility attribution?
- Or counterfactual reasoning about interventions?
- Or default inaction bias?

Relational reasoning:

- Reflects genuine care ethics commitments?
- Or in-group favoritism bias?
- Or reciprocity heuristics?

Future direction:

 doi:10.20944/preprints202602.1434.v1

- Process-tracing methods (think-aloud protocols, eye-tracking)
- Psychophysiological measures (emotional arousal during judgment)
- Computational modeling of decision processes
- Neuroscience approaches (fMRI studies of moral judgment)

4. Limited ecological validity of "pure" feature manipulations

We isolated features systematically (identifiability varied while holding stakeholder proximity constant, etc.). Real organizational decisions involve correlated features:

- Direct stakeholders are often identifiable
- Immediate consequences are often more certain
- Relational stakeholders often have longer history

Question: Do effects persist when features naturally covary?

Future direction:

- Conjoint analysis with realistic feature correlations
- Multi-level modeling of feature interactions
- Case studies examining how features combine in actual decisions

4.4.3. Future Research Directions

Beyond addressing limitations, several promising research directions emerge:

1. Optimal calibration studies

Research question: What combination of CR, AV, and SC produces best long-term organizational outcomes?

Approach:

- Longitudinal tracking of decision-makers◎variation profiles
- Measure downstream outcomes:
  - Stakeholder satisfaction
  - Decision quality (by independent evaluation)
  - Organizational culture/trust
  - Legal/compliance issues
- Test whether "Cluster 2" profile (CR = 0.28, AV = 0.31) predicts better outcomes

Hypothesis: Moderate CR with low AV maximizes stakeholder acceptance while minimizing arbitrary unfairness

2. Developmental trajectories

Research question: How does moral reasoning evolve with experience, training, and organizational socialization?

Approach:

- Compare novice vs. experienced managers
- Pre-post ethics training assessment
- Longitudinal panel following managers over 3-5 years
- Measure whether:
  - CR increases, decreases, or stabilizes
  - AV decreases (learning consistency)
  - Framework integration increases

Hypothesis: Experience → increased conditional consistency without decreased CR (learning appropriate framework selection)

3. Cultural variation in contextual patterns

Research question: Do non-Western cultural contexts show different contextual sensitivity patterns?

Approach:

- Replicate study in collectivist cultures (East Asia, Latin America)

 doi:10.20944/preprints202602.1434.v1

- Test whether:
  - Relational effects are stronger in collectivist cultures
  - Identifiability effects differ (named individuals vs. in-group/out-group)
  - Different frameworks dominate (harmony-focused vs. rights-focused)

Hypothesis: Collectivist cultures show higher baseline CR but lower AV (more systematic relational reasoning)

4. AI value alignment research

Research question: Can AI systems be trained to match human moral reasoning patterns including contextual sensitivity?

Approach:

- Fine-tune models on decisions from high-performing humans (Cluster 2 profile)
- Test whether fine-tuned models:
  - Reduce AV while maintaining moderate CR
  - Show systematic framework selection
  - Better match human variation patterns across temperature range

Hypothesis: Targeted training can achieve human-like calibration without temperature manipulation

5. Mechanism-focused experiments

Research question: What psychological processes mediate contextual effects?

Experiments:

A. Identifiability mechanism:

- Manipulate emotional content independently of identification
- Test mediation by empathy (self-report, psychophysiological)
- Distinguish affective vs. cognitive mechanisms

B. Relational mechanism:

- Manipulate relationship strength orthogonally to fairness considerations
- Include reciprocity measures
- Test whether effects persist when controlling for reciprocity

C. Action-omission mechanism:

- Manipulate causal language with constant outcomes
- Measure responsibility attribution explicitly
- Test whether making causation salient eliminates effect

6. Organizational field experiments

Research question: Do interventions targeting optimal variation profiles improve decision quality?

Interventions:

A. Decision support systems:

- Provide real-time feedback on CR/AV/SC metrics
- Flag unexplained variation for reflection
- Test whether feedback reduces AV without eliminating CR

B. Structured decision protocols:

- Require explicit framework identification
- Document contextual features considered
- Test whether documentation improves conditional consistency

C. AI-assisted decision-making:

- Deploy hybrid ensemble model (§4.3.2, Model 3)
- Compare decision quality to human-only or AI-only conditions
- Measure stakeholder acceptance across conditions

 doi:10.20944/preprints202602.1434.v1

7. Normative-empirical integration

Research question: Can empirical findings inform normative theory development?

Approach:

- Delphi study with moral philosophers
- Present empirical findings (e.g., "relational reasoning reduces AV")
- Ask: Does this evidence change your normative view of relational ethics?
- Iterative dialogue between empirical and normative researchers

Goal: Develop normatively defensible and empirically realistic theories of organizational ethics

8. Broader domain testing

Current study: Organizational ethics scenarios

Extensions:

A. Personal moral dilemmas:

- Classic trolley problems with contextual manipulations
- Test whether organizational effects generalize to personal ethics
- Compare professionals to general population

B. Political/policy decisions:

- Public officials facing resource allocation
- Test identifiability effects in policy contexts
- Compare elected officials to appointed administrators

C. Medical ethics:

- Physician decision-making with patient variation
- Test relational effects in doctor-patient relationships
- Compare to medical AI recommendation systems

D. Legal judgment:

- Judges/juries with defendant/victim variation
- Test whether legal training reduces contextual effects
- Compare to algorithmic sentencing recommendations

4.5. Conclusions and Recommendations

This study demonstrates that both humans and AI systems exhibit substantial contextual sensitivity in moral judgment, with 22-24% of variation attributable to features whose moral relevance is philosophically contested (FDR-corrected $q < .001$ for all primary effects). This creates a paradox: Identical empirical patterns support diametrically opposed normative conclusions depending on meta-ethical commitments.

Key conclusions:

1. **Contextual sensitivity is systematic, not random**

The finding that relational reasoning increases CR while decreasing AV (§3.5.4-3.5.5) contradicts the interpretation that context-sensitivity represents confused or inconsistent thinking. Instead, it appears to be one systematic moral consideration among others.

2. **Perfect consistency is neither achievable nor clearly desirable**

Even at very low AI temperature (0.3), contextual effects persist ($\eta^2p = 0.16$-0.18 for all four features, all $p < .001$, FDR-corrected $q < .001$). The "optimal" profile (Cluster 2) involves moderate CR (0.28), not zero. Organizations should target calibration, not elimination, of contextual sensitivity.

3. **Framework integration can be sophisticated or confused**

High integrators (Level 3-4) show both benefits (systematic framework selection, higher conditional consistency) and costs (elevated overall variation, higher AV). The key question is whether framework switching is principled (particularist interpretation) or arbitrary (principlist interpretation).

 doi:10.20944/preprints202602.1434.v1

4 Humans and AI differ primarily in relational reasoning

The 69% mediation by relational reasoning (indirect effect $\beta = 0.043$, 95% CI [0.028, 0.061], FDR-corrected $q < .001$) suggests this is the key dimension where human moral reasoning diverges from current AI systems. Whether this represents human superiority (particularist view) or human bias (principlist view) remains contested.

5. **Empirical findings constrain but cannot resolve normative debates**

We can measure contextual sensitivity and its correlates, but cannot determine whether it is morally appropriate. Normative argument remains essential. Our contribution is providing precise empirical benchmarks for theoretical debate.

Recommendations for practice:

For organizations:

1. **Acknowledge contextual sensitivity** rather than demanding impossible consistency
2. **Set empirical targets** (CR = 0.25-0.30, AV < 0.30, SC > 0.85) aligned with bounded particularism
3. **Audit variation components separately** (structural, contextual, arbitrary)
4. **Document framework selection criteria** to ensure principled rather than arbitrary context-sensitivity
5. **Configure AI systems intentionally** with temperature matching organizational meta-ethical commitments

For ethics training:

1. **Teach multiple frameworks** with explicit scope conditions
2. **Practice framework selection** systematically, not ad hoc
3. **Provide feedback on variation profiles** (individual CR/AV/SC metrics)
4. **Target optimal calibration** (Cluster 2 profile), not zero sensitivity
5. **Distinguish morally relevant from irrelevant context** explicitly

For AI governance:

1. Make temperature selection explicit governance decision
2. Deploy ensemble models combining different temperature settings
3. Require framework documentation in AI-generated recommendations
4. Monitor CR/AV/SC metrics as deployment KPIs
5. Preserve human authority for high-stakes decisions in contested domains

For future research:

1. **Field studies** of actual organizational decisions
2. **Cross-cultural replication** to test generalizability
3. **Mechanism experiments** to identify underlying processes
4. **Longitudinal tracking** of variation profile development
5. **Normative-empirical integration** through interdisciplinary dialogue

Final reflection:

The context sensitivity paradox reveals a fundamental challenge for organizational ethics: The same moral psychology that enables rich, contextually attuned judgment also creates systematic departures from principled consistency. Rather than viewing this as a defect to eliminate, we propose treating it as a design challenge: How can we preserve appropriate moral sensitivity to contextual features while minimizing arbitrary variation?

Our data suggest this is achievable. The "Balanced Sensitivity" profile (Cluster 2, representing 31% of humans and 48% of AI responses at temperature 0.7) demonstrates that moderate contextual responsiveness (CR = 0.28) can coexist with low arbitrary variation (AV = 0.31, lowest among clusters). This profile balances:

- Principlist concerns (structural consistency 86%, low arbitrary variation)
- Particularist insights (systematic sensitivity to relational and concrete contextual features)

 doi:10.20944/preprints202602.1434.v1

Whether this represents optimal moral reasoning or sophisticated bias remains philosophically contested. But it provides an empirical target for organizations seeking to navigate between the extremes of rigid principlism and unconstrained particularism.

The choice between these interpretations is not merely academic—it shapes ethics training, AI deployment, stakeholder communication, and organizational culture. Our hope is that by precisely mapping the landscape of moral variation, we enable more informed debate about where, and how much, context should matter in organizational ethics.

5. Conclusions

5.1. Summary of Contributions

This study makes four primary contributions:

1. Empirical: Comprehensive mapping of moral variation

We provide the first systematic measurement of how contextual features influence moral judgment in organizational contexts:

- **Overall variation:** Both humans (42%) and AI (41%) show substantial variation across scenario presentations
- **Component decomposition:** 85% structural consistency, 22-24% contextual responsiveness, 32-34% arbitrary variation
- **Feature effects:** Identifiability (OR=2.08), relational proximity (OR=7.89), temporal proximity (OR=1.52), relational context (OR=1.89) all produce large, robust effects
- **Human-AI similarity:** At deployment parameters (temperature 0.7), AI replicates human patterns (no significant Source × Feature interactions)

2. Theoretical: Precision on the consistency paradox

We make precise **which variation is normatively contested**:

- **85% structural consistency** is achievable (not controversial—should approach 100%)
- **32-34% arbitrary variation** is problematic (not controversial—should be minimized)
- **22-24% contextual responsiveness** is the **contested philosophical zone**:
    - Principlism: Bias requiring elimination
    - Particularism: Appropriate moral sensitivity

The paradox: Identical empirical patterns receive opposite normative interpretations depending on whether contextual features are deemed morally relevant.

3. Methodological: New calibration framework

We propose evaluating moral reasoning quality by **calibration** (systematic sensitivity relative to arbitrary noise) rather than raw consistency:

Metrics:

- **Structural Consistency (SC):** Agreement when irrelevant features vary
- **Contextual Responsiveness (CR):** Variation attributable to debatable features
- **Arbitrary Variation (AV):** Unexplained residual
- **Calibration:** CR/AV ratio (signal-to-noise)

Key finding: Optimal profiles (Clusters 2-3) show **moderate CR with low AV**, not **low CR** (Cluster 1 shows highest AV).

4. Practical: Implications for organizations and AI governance

We provide actionable recommendations:

For organizations:

- Aim for calibrated sensitivity (moderate CR, low AV), not maximum consistency
- Make contextual features explicit in decision processes
- Use domain-specific consistency standards
- Audit for systematic bias vs. appropriate sensitivity

 doi:10.20944/preprints202602.1434.v1

For AI development:

- Acknowledge that temperature/design choices embed normative commitments
- Develop explicit relational reasoning modules
- Target calibration (CR/AV > 0.8), not just consistency
- Report calibration metrics alongside performance metrics

For regulation:

- Specify calibration targets, not just "consistency"
- Audit for three components (SC, CR, AV) separately
- Set domain-specific standards
- Measure signal-to-noise, not just noise

5.2. The Irreducible Normative Question

Our central conclusion:
Empirical research can **constrain** but not **resolve** philosophical debates about moral reasoning.
We have established:

- ✓ Substantial contextual variation exists (22-24% of variance) in both humans and AI at typical deployment parameters (T=0.7)[1]
- ✓ Variation is systematic (framework-appropriate, reliable)
- ✓ Some variation is problematic (32-34% arbitrary)
- ✓ Moderate sensitivity outperforms zero sensitivity on reducing arbitrary variation
- ✓ Patterns are shared by humans and AI

We cannot establish:

- ✗ Whether contextual features **should** influence judgment
- ✗ What level of CR is "optimal" (depends on normative framework)
- ✗ How AI systems **should** be designed (depends on values)

The philosophical work ahead:
For principlists:

- Explain why zero sensitivity increases arbitrary variation
- Explain framework-appropriate patterns (if pure bias, shouldn't be framework-specific)
- Develop training that reduces arbitrary variation while pursuing consistency

For particularists:

- Explain 32-34% arbitrary variation (what accounts for unjustified inconsistency?)
- Provide principles for when context matters (avoid "anything goes" relativism)
- Distinguish good from bad particularism (focused vs. diffuse sensitivity)
- Address organizational implementation (how to scale case-by-case judgment)

For all:

- Engage with empirical constraints (theories must explain observed patterns)
- Specify which features are relevant and why
- Provide decision procedures for contested cases
- Test whether normative principles improve calibration empirically

[1] AI patterns observed at temperature=0.7, which was selected post-hoc to match human total variation levels. At lower temperatures (T=0.3), AI shows significantly less contextual responsiveness (CR=0.12 vs. human 0.27, p<.001); at higher temperatures (T=1.0), AI shows similar CR (0.28) but with elevated arbitrary variation (AV=0.49 vs. human 0.33, p<.001) and degraded coherence. Human-AI similarity is thus parameter-dependent, not inherent to AI reasoning architecture. However, the existence and direction of contextual effects (identifiability OR>1.3, relational OR>1.4, temporal OR>1.2, action-omission d>0.5) persist across all temperatures (T=0.3 to T=1.0, all p<.001), supporting robustness of core findings independent of calibration choices.

 doi:10.20944/preprints202602.1434.v1

A Critical Caveat: The Temperature Parameter as Normative Choice

Before proceeding to normative implications, we must acknowledge a fundamental limitation: **our human-AI comparisons are entangled with temperature calibration choices**.

The circularity:

1. We selected temperature T=0.7 because it produced human-like total variation (0.41 vs. 0.42, p=.56)
2. At T=0.7, we found that AI exhibits human-like contextual sensitivity patterns
3. We cannot therefore claim AI "inherently" reasons like humans

What this means for interpretation:

Claims we CAN make:

- ✓ At deployment parameters that match human total variation, AI exhibits similar contextual responsiveness (CR ≈ 0.24 for both)
- ✓ Contextual effects exist across the full temperature range (T=0.3 to T=1.0, all p<.001)
- ✓ Temperature is a design choice that embeds normative commitments about desired reasoning patterns
- ✓ Organizations can calibrate AI to target specific variation profiles (principlist T=0.3, moderate T=0.7, particularist T=1.0)

Claims we CANNOT make:

- ✗ AI reasoning is fundamentally similar to human reasoning
- ✗ AI would exhibit human-like patterns at "default" or "optimal" settings absent calibration
- ✗ Human-AI convergence is independent of parameter selection
- ✗ AI "naturally" achieves the balanced profile (Cluster 2) more frequently than humans

Why temperature matters normatively:

The temperature parameter controls the CR/AV trade-off:

- **Low temperature (T=0.3)**: CR=0.12, AV=0.41, coherence=99.6%
  - Principlist-friendly: minimal contextual sensitivity
  - But paradoxically higher arbitrary variation than moderate temperature
- **Moderate temperature (T=0.7)**: CR=0.24, AV=0.34, coherence=97.2%
  - Human-matched: similar CR and AV to human average
  - Balanced calibration between sensitivity and noise
- **High temperature (T=1.0)**: CR=0.28, AV=0.49, coherence=91.6%
  - Particularist-friendly: high contextual sensitivity
  - But excessive arbitrary variation and coherence degradation

Selecting temperature is thus **implicitly selecting a normative position** on the appropriate balance between:

- Contextual responsiveness (particularist value)
- Consistency (principlist value)
- Coherence (rational discourse requirement)

Implications for our "bounded particularism" proposal:

Our recommendation of moderate CR (0.25-0.30) and low AV (<0.30) is supported by:

- ✓ Human-only clustering showing this profile minimizes arbitrary variation (Cluster H2: AV=0.29)
- ✓ Theoretical argument that neither extreme (pure principlism nor unbounded particularism) achieves low AV
- ✓ Practical considerations about realistic targets for ethics training

But the recommendation is **not** supported by:

 doi:10.20944/preprints202602.1434.v1

- ✗ Observation that AI "naturally" achieves this profile (AI frequency in Cluster 2 is temperature-dependent)
- ✗ Claim that this profile represents AI "reasoning" superiority

Transparency in reporting:

Throughout this paper, when we claim "AI exhibits pattern X," readers should understand this as:

- "AI at temperature T=0.7 (selected to match human variation) exhibits pattern X"
- NOT "AI inherently exhibits pattern X"

This caveat is crucial for:

- Practitioners deciding how to configure AI systems (temperature is a choice, not a given)
- Researchers interpreting human-AI comparisons (similarity is parameter-dependent)
- Philosophers evaluating whether AI "reasoning" provides evidence for moral theories (it doesn't - it provides evidence about what patterns emerge at different calibration settings)

The fundamental insight:

AI temperature selection is not a technical detail—it is a disguised normative commitment about the appropriate balance between contextual sensitivity, consistency, and decision coherence. Making this choice explicit transforms AI ethics from technical optimization to value-laden calibration.

Our study provides empirical mapping of the CR/AV/coherence trade-offs across temperature settings, but selecting the "optimal" point on this trade-off curve remains an irreducibly normative question that AI system designers must answer based on philosophical commitments and organizational values.

5.3. Future Outlook

We envision three trajectories:

Trajectory 1: Continued debate without resolution

The principlism-particularism dispute remains fundamentally irresolvable. Empirical evidence accumulates but philosophical positions remain entrenched. Organizations and AI developers make **implicit choices** that favor one view without acknowledging contestation.

Risk: Hidden normative commitments in "neutral" systems.

Trajectory 2: Pluralist framework emerges

Recognition that **different contexts warrant different standards**:

- Rule-based domains → principlism appropriate (maximize consistency)
- Relationship domains → particularism appropriate (calibrated sensitivity)
- Justice domains → framework pluralism (multiple legitimate principles)

Organizations and AI explicitly select calibration targets based on domain. Transparency about normative commitments replaces false neutrality.

Advantage: Pragmatic, context-sensitive approach.

Trajectory 3: Empirically-grounded convergence

Accumulating evidence about **which patterns predict good outcomes** (decision quality, stakeholder satisfaction, long-term organizational success) empirically constrains philosophical debate.

Example:

- If high CR with low AV predicts better organizational outcomes, particularism vindicated
- If low CR predicts better outcomes, principlism vindicated
- If outcome-dependence itself emerges (context-specific optimality), pluralism vindicated

Limitation: Requires agreement on **outcome metrics** (itself normatively contested).

Our hope:

 doi:10.20944/preprints202602.1434.v1

This study provides **empirical scaffolding** for normative debate—precise measurements of what patterns exist, how systematic they are, where philosophical dispute is located.

The path forward requires **integration** of:

- Rigorous empirical measurement (which patterns exist?)
- Careful philosophical analysis (which patterns should exist?)
- Practical experimentation (which approaches work in organizations?)
- Iterative refinement (revise both principles and practices based on evidence)

The question "How much moral variation is too much?" has no purely empirical answer.

But empirical research can make the question **precise**, identify **where philosophical work is needed**, and test **whether proposed principles improve outcomes**.

In this spirit, we offer our findings:

Not as resolution, but as **clarification** of the empirical landscape on which normative arguments must stand.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

References

Aristotle. (1999). *Nicomachean ethics* (2nd ed., T. Irwin, Trans.). Hackett Publishing. (Original work published ca. 350 BCE)

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... & Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Baron, J. (1992). The effect of normative beliefs on anticipated emotions. *Journal of Personality and Social Psychology*, *63*(2), 320-330.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74-85.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*(1), 1-16.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*(2), 381-417.

Bazerman, M. H., & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what's right and what to do about it*. Princeton University Press.

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

Bentham, J. (1996). *An introduction to the principles of morals and legislation*. Oxford University Press. (Original work published 1789)

Bloom, P. (2017). *Against empathy: The case for rational compassion*. Ecco/HarperCollins.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*, 4299-4307.

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 39-67.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273-292.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082-1089.

Dancy, J. (1993). *Moral reasons*. Blackwell.

 doi:10.20944/preprints202602.1434.v1

Dancy, J. (2004). *Ethics without principles*. Oxford University Press.

Dancy, J. (2013). Moral particularism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition). https://plato.stanford.edu/archives/win2013/entries/moral-particularism/

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880-896.

Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review*, *20*(1), 65-91.

Dworkin, R. (1977). *Taking rights seriously*. Harvard University Press.

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5-15.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42.

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman.

Freeman, R. E., Harrison, J. S., Wicks, A. C., Parmar, B. L., & De Colle, S. (2010). *Stakeholder theory: The state of the art*. Cambridge University Press.

Fried, C. (1978). *Right and wrong*. Harvard University Press.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411-437.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., ... & Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

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

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

Goodman, N. (1954). *Fact, fiction, and forecast*. Harvard University Press.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55-130.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366-385.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322-323.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

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

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

Haidt, J., & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology*, *26*(2), 201-218.

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology*, *31*(1), 191-221.

Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford University Press.

Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.

Held, V. (2006). *The ethics of care: Personal, political, and global*. Oxford University Press.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hooker, B., & Little, M. (Eds.). (2000). *Moral particularism*. Oxford University Press.

Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: On the affective psychology of value. *Journal of Experimental Psychology: General*, *133*(1), 23-30.

Hume, D. (1978). *A treatise of human nature* (2nd ed., L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford University Press. (Original work published 1739-1740)

Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, *14*(3), 235-257.

 doi:10.20944/preprints202602.1434.v1

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.

Kagan, S. (1989). *The limits of morality*. Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291.

Kant, I. (1993). *Grounding for the metaphysics of morals* (3rd ed., J. W. Ellington, Trans.). Hackett Publishing. (Original work published 1785)

Kant, I. (1996). *Practical philosophy* (M. J. Gregor, Trans. & Ed.). Cambridge University Press.

Kohlberg, L. (1984). *Essays on moral development: Vol. 2. The psychology of moral development*. Harper & Row.

Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.

Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingis, Trans.). Duquesne University Press.

Levinas, E. (1985). *Ethics and infinity: Conversations with Philippe Nemo* (R. A. Cohen, Trans.). Duquesne University Press.

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, *25*(5), 661-671.

Lickona, T. (1991). *Educating for character: How our schools can teach respect and responsibility*. Bantam Books.

Little, M. O. (2000). Moral generalities revisited. In B. Hooker & M. Little (Eds.), *Moral particularism* (pp. 276-304). Oxford University Press.

MacIntyre, A. (1981). *After virtue: A study in moral theory*. University of Notre Dame Press.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224-253.

McCloskey, H. J. (1957). An examination of restricted utilitarianism. *Philosophical Review*, *66*(4), 466-485.

McDowell, J. (1979). Virtue and reason. *The Monist*, *62*(3), 331-350.

McDowell, J. (1998). *Mind, value, and reality*. Harvard University Press.

McNaughton, D. (1988). *Moral vision: An introduction to ethics*. Blackwell.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143-152.

Mill, J. S. (1998). *Utilitarianism* (R. Crisp, Ed.). Oxford University Press. (Original work published 1861)

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306-1315.

Moore, G. E. (1903). *Principia ethica*. Cambridge University Press.

Nagel, T. (1970). *The possibility of altruism*. Princeton University Press.

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530-542.

Noddings, N. (1984). *Caring: A feminine approach to ethics and moral education*. University of California Press.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

Nussbaum, M. C. (1990). *Love's knowledge: Essays on philosophy and literature*. Oxford University Press.

Nussbaum, M. C. (2001). *Upheavals of thought: The intelligence of emotions*. Cambridge University Press.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Parfit, D. (2011). *On what matters* (Vols. 1-2). Oxford University Press.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*(3), 145-171.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology*, *64*(3), 467-478.

Pettit, P., & Brennan, G. (1986). Restrictive consequentialism. *Australasian Journal of Philosophy*, *64*(4), 438-455.

 doi:10.20944/preprints202602.1434.v1

Piaget, J. (1965). *The moral judgment of the child* (M. Gabain, Trans.). Free Press. (Original work published 1932)

Portmore, D. W. (2011). *Commonsense consequentialism: Wherein morality meets rationality*. Oxford University Press.

Prinz, J. J. (2007). *The emotional construction of morals*. Oxford University Press.

Railton, P. (1984). Alienation, consequentialism, and the demands of morality. *Philosophy & Public Affairs*, *13*(2), 134-171.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rawls, J. (1993). *Political liberalism*. Columbia University Press.

Raz, J. (2000). The truth in particularism. In B. Hooker & M. Little (Eds.), *Moral particularism* (pp. 48-78). Oxford University Press.

Rest, J. R. (1979). *Development in judging moral issues*. University of Minnesota Press.

Robinson, P. H., & Darley, J. M. (1995). *Justice, liability, and blame: Community views and the criminal law*. Westview Press.

Ross, W. D. (1930). *The right and the good*. Oxford University Press.

Sandel, M. J. (1982). *Liberalism and the limits of justice*. Cambridge University Press.

Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32-70.

Scherrer, N., Shi, C., Feder, A., & Blei, D. (2024). Evaluating the moral beliefs encoded in LLMs. *Advances in Neural Information Processing Systems*, *37*.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1096-1109.

Sen, A. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, *6*(4), 317-344.

Sen, A. (2009). *The idea of justice*. Harvard University Press.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119-169). Routledge.

Sidgwick, H. (1907). *The methods of ethics* (7th ed.). Macmillan.

Simmons, G. (2023). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2309.12349*.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, *1*(3), 229-243.

Singer, P. (1979). *Practical ethics*. Cambridge University Press.

Singer, P. (2009). *The life you can save: Acting now to end world poverty*. Random House.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3-22.

Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making*, *2*(2), 79-95.

Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, *26*(1), 5-16.

Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, *102*(2), 143-153.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., ... & Choi, Y. (2024). Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22000-22310.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76-105.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645-665.

Starmans, C., & Bloom, P. (2016). When the spirit is willing, but the flesh is weak: Developmental differences in judgments about inner moral conflict. *Psychological Science*, *27*(11), 1498-1506.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531-542.

 doi:10.20944/preprints202602.1434.v1

Sunstein, C. R. (2019). *How change happens*. MIT Press.

Talisse, R. B., & Aikin, S. F. (2008). *Pragmatism: A guide for the perplexed*. Continuum.

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., ... & Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Social Cognitive and Affective Neuroscience*, *7*(3), 282-288.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*(7), 320-324.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*(6), 1395-1415.

Thomson, J. J. (1990). *The realm of rights*. Harvard University Press.

Tronto, J. C. (1993). *Moral boundaries: A political argument for an ethic of care*. Routledge.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*(6), 476-491.

Unger, P. (1996). *Living high and letting die: Our illusion of innocence*. Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998-6008.

Waldron, J. (1987). Theoretical foundations of liberalism. *The Philosophical Quarterly*, *37*(147), 127-150.

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

Walzer, M. (1994). *Thick and thin: Moral argument at home and abroad*. University of Notre Dame Press.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.

Weidman, A. C., Sowden, W. J., Berg, M. K., & Kross, E. (2020). Punish or protect? How close relationships shape responses to moral violations. *Personality and Social Psychology Bulletin*, *46*(5), 693-708.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780-784.

Williams, B. (1973). A critique of utilitarianism. In J. J. C. Smart & B. Williams, *Utilitarianism: For and against* (pp. 77-150). Cambridge University Press.

Williams, B. (1985). *Ethics and the limits of philosophy*. Harvard University Press.

Wistrich, A. J., Guthrie, C., & Rachlinski, J. J. (2005). Can judges ignore inadmissible information? The difficulty of deliberately disregarding. *University of Pennsylvania Law Review*, *153*(4), 1251-1345.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202-214.

Zheng, R., Consoli, S., & Zhao, L. (2023). Large language models for ethics: A systematic literature review. *arXiv preprint arXiv:2308.12711*.