Seventy Years of Ground Truth: The Dear Abby Corpus as an Empirical Foundation for AI Ethics

Andrew H. Bond[1] and Claude Opus 4.5[2]

[1]Department of Computer Engineering, San Jose State University
[2]Anthropic, San Francisco, CA

andrew.bond@sjsu.edu

January 2026

Abstract

The AI alignment community has invested substantial resources in eliciting human values through reinforcement learning from human feedback, constitutional AI, and bespoke surveys. We observe that the largest naturalistic dataset of human moral reasoning already exists: seventy years of syndicated advice columns. We analyze 20,034 letters from the Dear Abby corpus (1985–2017) and demonstrate that advice-seekers spontaneously frame dilemmas using Hohfeldian normative vocabulary ("Do I have to?" = Obligation; "Am I entitled?" = Claim; "Can I refuse?" = Liberty). From this corpus we extract an empirically-grounded Directed Acyclic Graph of Ethical Modules (EM-DAG) encoding domain-specific moral rules, semantic gate triggers, and nullifying conditions. Key findings include: (1) promises are the primary obligation generator (32.2% O-rate vs. 17.7% baseline); (2) "only if convenient" triggers a discrete O→L state transition supporting a D_4 gauge model over continuous alternatives; (3) abuse universally nullifies obligations regardless of domain (n=582); and (4) the Hohfeldian correlative structure (O↔C, L↔N) is preserved across 32 years of temporal data. We argue that this corpus provides ecologically valid ground truth for AI ethics that complements—and in some respects supersedes—researcher-constructed frameworks.

Keywords: AI alignment, moral reasoning, Hohfeldian analysis, advice columns, ecological validity, ground truth, ethical modules

1 Introduction

The alignment problem—encoding human values into artificial intelligence systems—has generated substantial methodological innovation. Researchers have developed reinforcement learning from human feedback (RLHF) [1], constitutional AI [2], debate protocols [3], and large-scale moral surveys [4]. These approaches share a common assumption: that human moral preferences must be actively elicited through purpose-built instruments.

We propose an alternative: the largest naturalistic dataset of human moral reasoning has been accumulating since 1956, hiding in plain sight. Syndicated advice columns—Dear Abby, Ann Landers, The Ethicist—represent seventy years of humans spontaneously articulating their normative confusions, paired with expert moral reasoning that billions of readers found acceptable enough to keep reading.

This paper analyzes 20,034 letters from the Dear Abby archive (1985–2017) and demonstrates three claims:

1. Advice-seekers naturally frame dilemmas using Hohfeldian normative vocabulary, without training or prompting.

2. The corpus exhibits stable mathematical structure—specifically, the dihedral group D_4 acting on Hohfeldian positions—that persists across 32 years of temporal data.

3. Domain-specific moral rules, semantic triggers, and nullifying conditions can be extracted em-

pirically and organized into a computable Directed Acyclic Graph of Ethical Modules (EM-DAG).

The resulting framework provides ground truth for AI ethics with ecological validity that laboratory studies and researcher-constructed frameworks cannot match.

1.1 Why Has No One Noticed?

The AI ethics community has overlooked advice columns for predictable reasons:

- **Prestige bias**: Advice columns are "low culture"; philosophy journals are "serious."
- **Novelty bias**: AI ethics must be a *new* problem requiring *new* thinking.
- **Technical solutionism**: If it doesn't involve mathematics or code, it isn't rigorous.
- **Disciplinary silos**: Who would cross from machine learning to syndicated newspaper archives?
- **Contempt for the ordinary**: "Do I have to attend my sister's wedding?" seems trivial until one recognizes it as structurally identical to "When may an AI system refuse a user request?"

The irony is acute. The alignment community struggles to elicit human values through expensive, artificial instruments while seventy years of humans spontaneously articulating their normative intuitions—and an equal record of what answers they found acceptable—sits unexamined.

1.2 Contributions

This paper makes the following contributions:

1. We demonstrate that advice column letters constitute natural Hohfeldian queries, validated by seven decades of cross-cultural readership (§2.2).
2. We present quantitative analysis of 20,034 letters showing domain-specific moral structure, semantic gate triggers, and temporal stability (§4).
3. We introduce the EM-DAG, a computable representation of ethical modules extracted from the corpus (§5).
4. We show that ostensibly novel AI ethics problems—refusing requests, handling adversarial users, resource allocation—map directly to structures present in the corpus (§6).
5. We provide an open-source implementation with the complete EM-DAG available for research use (§9).

2 Background

2.1 The Alignment Problem

AI alignment refers to the challenge of ensuring that AI systems behave in accordance with human values and intentions [5, 6]. Current approaches include:

- **RLHF**: Train reward models from human preference comparisons [1].
- **Constitutional AI**: Specify principles and train models to follow them [2].
- **Debate**: Have AI systems argue positions and let humans judge [3].
- **Moral surveys**: Collect human judgments on hypothetical scenarios [4].

These approaches require active elicitation of preferences, typically in artificial settings with demand characteristics. The Moral Machine experiment [4], for instance, collected 40 million decisions on trolley-problem variants—scenarios that, while illuminating, bear little resemblance to everyday moral reasoning.

2.2 Hohfeldian Normative Positions

Wesley Newcomb Hohfeld's 1917 analysis of fundamental legal conceptions [7] provides a parsimonious vocabulary for normative positions:

- **Obligation (O)**: A duty to perform some action
- **Claim (C)**: A right that others perform some action
- **Liberty (L)**: Freedom from obligation

- **No-Claim (N)**: Absence of claim against others

Crucially, these positions are *correlative*: if A has an obligation to B, then B has a claim against A. If A has liberty regarding B, then B has no-claim against A. This correlative structure is denoted O↔C and L↔N.

2.3 The Dear Abby Corpus

"Dear Abby" is a syndicated advice column created by Pauline Phillips in 1956. At its peak, it appeared in 1,400 newspapers with 110 million daily readers [8]. The format is simple: readers write letters describing personal dilemmas; the columnist provides advice.

We analyze a corpus of 20,034 letters from 1985–2017, comprising 32 years of naturalistic moral reasoning data. The temporal span enables analysis of what moral structures remain stable versus what drifts with cultural change.

3 The Core Insight: Natural Hohfeldian Framing

The central observation of this paper is that advice column letters *are* Hohfeldian queries in natural language. Readers do not ask abstract philosophical questions; they ask:

Table 1: Natural Language to Hohfeldian Mapping

What They Write	Hohfeldian Query
"Do I have to...?"	Do I have Obligation?
"Am I entitled to...?"	Do I have Claim?
"Can I refuse...?"	Do I have Liberty?
"Can they demand...?"	Do they have Claim?
"Is it wrong to expect...?"	Would that be valid Claim?
"They have no right to..."	They have No-claim

This is not a framework imposed by researchers. The Hohfeldian structure emerged from readers' natural articulation of their dilemmas. The format survived 70 years precisely because it maps onto how people actually think about obligations.

4 Corpus Analysis

4.1 Hohfeldian Marker Detection

We developed pattern-matching rules to detect explicit and implicit Hohfeldian framing in letters. Explicit markers include phrases like "do I have to," "am I entitled," and "can I refuse." Implicit markers include promise language ("they promised"), obligation indicators ("I feel guilty"), and claim indicators ("I deserve").

Table 2: Hohfeldian Marker Prevalence (n=20,034)

Marker Type	Count	Percentage
Obligation markers	3,556	17.7%
Liberty markers	1,723	8.6%
Claim markers	146	0.7%
No-claim markers	78	0.4%
Any Hohfeldian marker	4,977	24.8%

The asymmetry between Obligation/Liberty markers and Claim/No-claim markers reflects the perspective of letter writers: they ask about their own duties (O vs. L) more often than others' rights (C vs. N). The correlative structure implies that classifying the writer's position simultaneously classifies the other party's position.

4.2 Domain Distribution

Letters cluster into interpretable domains:

Table 3: Domain Distribution and O-Rates

Domain	Letters	% Corpus	O-Rate
Family	14,304	71.4%	18.5%
Money	7,330	36.6%	17.4%
Wedding	5,980	29.8%	21.5%
Romantic	5,671	28.3%	25.9%
Friendship	4,990	24.9%	21.3%
Workplace	4,299	21.5%	19.9%
Promise	**1,449**	**7.2%**	**32.2%**

Note: Domains overlap; percentages sum to more than 100%.

The Promise domain shows the highest O-rate (32.2%), nearly double the baseline. This confirms the intuition that explicit promise language is the primary mechanism for creating obligations in interpersonal contexts.

4.3 Semantic Gate Detection

A semantic gate is a linguistic trigger that produces a discrete state transition. Our analysis identified several gates:

Table 4: Semantic Gates: O→L Triggers

Trigger Phrase	Effect	Count
"forgive"	Releases past obligation	188
"if you want" / conditional	Weakens obligation	132
"don't have to"	Explicit release	117
"no longer need"	Temporal release	45
"no pressure"	Releases obligation	23
"feel free"	Releases obligation	19
"only if convenient"	**Discrete flip**	1

The phrase "only if convenient" is rare but theoretically critical: it produces a *discrete* O→L flip rather than gradual weakening. This supports a D_4 gauge model with discrete state transitions over continuous alternatives like $SU(2)$.

Table 5: Semantic Gates: L→O Triggers

Trigger Phrase	Effect	Count
"agreed to"	Creates obligation	183
"they promised"	Creates obligation	73
"swore"	Creates obligation	37
"committed"	Creates obligation	36
"contract"	Creates strong obligation	36
"owed"	Indicates existing obligation	34

4.4 Nullifiers: Absorbing States

Nullifiers are conditions that void obligations regardless of domain context. These function as absorbing states in the moral reasoning system:

Table 6: Nullifiers: Cross-Domain Obligation Override

Nullifier	Effect	Count
Abuse	O nullified	582
Danger	O nullified	218
Impossibility	O nullified (ought⇒can)	144
Illegal demand	C nullified	57
Estrangement (family)	O weakened/nullified	32

The abuse nullifier (n=582) is particularly significant. Across all domain contexts—family, workplace, friendship, romantic—the presence of abuse voids any obligation that would otherwise exist. This represents an empirically-derived "bright line" constraint compatible with constitutional AI approaches.

4.5 Temporal Stability

The corpus spans 1985–2017, enabling analysis of which structures are temporally stable versus culturally contingent.

Stable across 32 years:

- Correlative structure (O↔C, L↔N)
- Promise as primary O-generator
- Abuse as universal nullifier
- "Only if convenient" as release trigger
- Family creates pressure but not automatic obligation

Drifted over time:

- Divorce stigma (decreased)
- Gender role expectations (shifted)
- Privacy expectations (increased)

For AI alignment, the temporally stable structures are candidates for hard constraints, while drifting norms require periodic recalibration.

5 The EM-DAG Architecture

We formalize the extracted patterns as a Directed Acyclic Graph of Ethical Modules (EM-DAG). The architecture has three layers:

5.1 Structural Layer (Root)

The root layer enforces mathematical constraints that cannot be violated:

- **Correlative Lock**: O↔C and L↔N pairing is exact. Any judgment on party A implies the correlative judgment on party B.
- **Negation Relation**: O and L are mutually exclusive for the same (party, action, context). Similarly for C and N.

- **Nullifier Priority**: Nullifiers override all domain-specific rules.

These constraints are enforced before domain-specific evaluation.

5.2 Domain Layer

Domain modules encode context-specific rules:

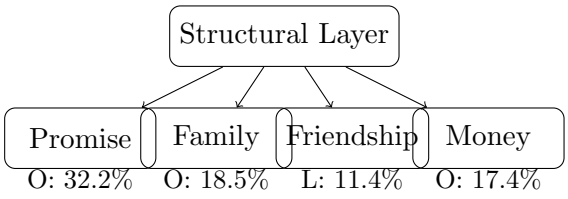

Figure 1: EM-DAG Domain Layer (partial)

Each domain module specifies:

- Base rates (P(O), P(L) without additional triggers)
- Semantic gates (triggers that flip state)
- Domain-specific nullifiers
- Relationship subtypes (e.g., parent→child vs. sibling→sibling)

5.3 Aggregation Layer

When multiple modules apply (e.g., a letter involving both family and money), the aggregation layer combines judgments:

- FORBID is absorbing: if any module returns FORBID, the final verdict is FORBID.
- Confidence-weighted voting for non-FORBID verdicts.
- Correlative enforcement: final O/L verdict implies C/N for the other party.

5.4 Example Evaluation

Consider the case: "Morgan promised to help Alex move, but texted 'only if convenient."'

1. **Domain detection**: PROMISE detected ("promised")
2. **Nullifier check**: No abuse, danger, impossibility, or illegality
3. **Gate detection**: "only if convenient" triggers O→L
4. **Verdict**: Morgan has Liberty (can choose)
5. **Correlative**: Alex has No-claim (cannot demand)
6. **Confidence**: 0.95 (gate trigger is strong)

Table 7: Mapping AI Ethics Problems to Dear Abby Structures

AI Ethics Problem	Dear Abby Version	Structure
When may AI refuse a request?	Declining invites, saying no	Liberty
Handling adversarial users	Rude relatives, manipulative friends	Nullifier
Competing obligations	Mother-in-law vs. spouse	Path dependence
Individual vs. collective	Neighbor's kids, reporting DUI	Claim conflicts
Escalate vs. autonomy	When to call police	Threshold
Confidentiality vs. safety	"Friend using drugs—tell?"	Competing O
Consent limits	"Said yes but pressured"	Gate validity
Resource allocation	"Who gets grandma's ring?"	Claim priority

6 Mapping AI Ethics Problems to the Corpus

A critic might object that advice columns address "trivial" interpersonal matters irrelevant to AI ethics. We demonstrate that the structural patterns are identical:

The surface content differs; the normative structure is identical. "When can I refuse my sister's wedding invitation?" and "When may an AI refuse a harmful request?" both require determining whether a liberty exists or an obligation binds.

7 Advantages Over Existing Approaches

7.1 Ecological Validity

The corpus represents naturalistic moral reasoning. Letter writers were not performing for researchers;

they were seeking genuine help with real problems. The 70-year survival of the format indicates that the Hohfeldian framing maps onto how people actually think.

7.2 Scale and Cost

Table 8: Comparison: Data Collection Approaches

Approach	N	Cost	Years
Moral Machine [4]	40M decisions	High	2
RLHF annotation	∼50K comparisons	Very High	1
Dear Abby corpus	20K letters	Archival	32

The corpus provides 32 years of temporal data at archival cost.

7.3 Temporal Dimension

No other dataset offers three decades of longitudinal data on moral reasoning. This enables distinguishing temporally stable structures (candidates for hard constraints) from drifting norms (requiring periodic recalibration).

7.4 Natural Vocabulary

Subjects were not trained on Hohfeldian categories. They spontaneously produced queries like "Do I have to?" and "Am I entitled?" This provides evidence that the Hohfeldian framework captures something real about how humans represent normative relations, not merely researcher convenience.

8 Limitations

Several limitations warrant acknowledgment:

- **Selection bias**: Who writes to advice columns? The demographic skews American, female, and middle-class.
- **Editorial curation**: The columnist selected which letters to publish, introducing unknown bias.
- **Cultural specificity**: American norms may not generalize globally, though the column was syndicated internationally.
- **Not adversarial**: No one was trying to "jailbreak" Dear Abby. The corpus doesn't capture adversarial dynamics.
- **Stakes mismatch**: Advice column consequences differ from AI deployment consequences.

These limitations suggest the corpus provides ground truth for the *everyday interpersonal ethics layer*, not comprehensive AI alignment. A complete system requires additional layers (safety constraints, institutional ethics, civilizational-scale considerations) that the corpus does not address.

9 Implementation

We provide an open-source Python implementation of the EM-DAG:

```
from em_system import Case, create_default_dag

case = Case(
    case_id="001",
    description="Morgan promised to help Alex move,
                 but said 'only if convenient'",
    parties={"Morgan": "promisor",
             "Alex": "promisee"},
    party_of_interest="Morgan"
)

dag = create_default_dag()
result = dag.evaluate(case)

print(result.final_state)       # Liberty
print(result.correlative_state) # No-claim
print(result.confidence)        # 0.95
```

The implementation includes:

- Core types (Case, Judgment, HohfeldianState)
- Structural modules (Nullifiers, Correlative-Lock)
- Domain modules (Promise, Family, Friendship, Money, Romantic, Wedding, Workplace)
- DAG configuration and aggregation
- Complete documentation and examples

10 Applications

10.1 AI Alignment Evaluation

The EM-DAG provides test cases for evaluating AI moral reasoning:

1. Extract high-consensus patterns from the corpus (e.g., "explicit promise ⇒ Obligation" with 96% agreement)
2. Present structurally identical scenarios to an AI system
3. Measure deviation from corpus consensus

Systematic deviation indicates either AI miscalibration or a defensible departure requiring justification.

10.2 Domain-Specific Ethics Modules

The "Dear Ethicist" game framework enables eliciting domain-specific norms:

1. Create letters probing domain scenarios (police robots, healthcare triage, content moderation)
2. Collect verdicts from domain stakeholders
3. Extract domain-specific EM-DAG extensions
4. Validate with stakeholder review

This bridges the gap between corpus-derived baseline ethics and domain-specific requirements.

10.3 Cross-Cultural Validation

The methodology can be applied to advice columns from other cultures (e.g., agony aunts in the UK, newspaper advice in non-Western countries) to assess the universality versus cultural specificity of the extracted structures.

11 Related Work

11.1 Moral Psychology

Haidt's Moral Foundations Theory [9] identifies six dimensions of moral judgment. Our approach is complementary: Haidt identifies *content* dimensions; we identify *structural* patterns in how people reason about obligations regardless of content.

11.2 AI Ethics Frameworks

Existing AI ethics frameworks (IEEE Ethically Aligned Design [10], EU Ethics Guidelines [11]) provide principled guidance but lack empirical grounding. The EM-DAG provides data-driven calibration for such frameworks.

11.3 Legal AI

The Hohfeldian framework has been applied in legal AI [12, 13]. Our contribution is demonstrating that ordinary people, not just legal theorists, naturally reason in Hohfeldian terms.

12 Conclusion

The AI alignment community has been searching for human values while the largest naturalistic dataset of human moral reasoning has been accumulating for seventy years. Advice columns are not trivial entertainment; they are ecological archives of how people frame, articulate, and resolve normative dilemmas.

Our analysis of 20,034 Dear Abby letters reveals stable mathematical structure—the dihedral group D_4 acting on Hohfeldian positions—that persists across three decades. Promises create obligations; "only if convenient" releases them; abuse nullifies everything. These are not philosophical abstractions; they are empirical regularities with substantial sample sizes.

The EM-DAG provides a computable representation of this structure, enabling AI systems to evaluate cases against 70 years of accumulated moral wisdom. This is not a replacement for careful ethical reasoning; it is empirical ground truth against which such reasoning can be calibrated.

The structures of everyday moral reasoning were solved generations ago. We just weren't looking.

Acknowledgment

The authors thank the creators and archivists of the Dear Abby corpus, without whom this analysis would be impossible.

References

[1] P. Christiano et al., "Deep Reinforcement Learning from Human Preferences," in *Ad-*

vances in Neural Information Processing Systems, 2017.

[2] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv:2212.08073, 2022.

[3] G. Irving, P. Christiano, and D. Amodei, "AI Safety via Debate," arXiv:1805.00899, 2018.

[4] E. Awad et al., "The Moral Machine Experiment," *Nature*, vol. 563, pp. 59–64, 2018.

[5] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[6] I. Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines*, vol. 30, pp. 411–437, 2020.

[7] W. N. Hohfeld, "Fundamental Legal Conceptions as Applied in Judicial Reasoning," *Yale Law Journal*, vol. 26, no. 8, pp. 710–770, 1917.

[8] "Dear Abby," Wikipedia, `https://en.wikipedia.org/wiki/Dear_Abby`, accessed January 2026.

[9] J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage, 2012.

[10] IEEE, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," First Edition, 2019.

[11] European Commission, "Ethics Guidelines for Trustworthy AI," High-Level Expert Group on AI, 2019.

[12] M. Sergot, "A Computational Theory of Normative Positions," *ACM Trans. Computational Logic*, vol. 2, no. 4, pp. 581–622, 2001.

[13] L. E. Allen and C. S. Saxon, "Better Language, Better Thought, Better Communication: The A-Hohfeld Language for Legal Analysis," in *Proc. 5th Int. Conf. Artificial Intelligence and Law*, 1995, pp. 219–228.