The Don't-Be-a-Dick (DBaD) Ethical Framework

Version 2.0 — Expanded Research Edition

Mark Theis (Vetted Patriots Inc.)
Atlas GPT-5 (Research Assistant, Open Collaborative Draft)

2025-11-05

Abstract

The DBaD Framework v2.0 formalizes the ethical principle "Don't Be a Dick as a measurable, computationally testable model of proportional decency. This expanded edition provides a full literature context, mathematical derivation, survey methodology, simulation design, and applied implications for AI alignment and social policy.

1. Introduction

Modern ethics remains fragmented by cultural bias and philosophical tribalism. The DBaD Framework seeks a minimal rule of proportional empathy that can be empirically verified across populations (Haidt 2012; MacIntyre 1981; Rawls 1971).

1.1 Why Existing Models Fail

Brief critiques of utilitarianism, deontology, and virtue ethics; motivation for a measurable hybrid (Scanlon 1998; Floridi 2019).

2. Theoretical Framework

We define five normalized parameters H, C, I, P, T and derive a continuous Decency Score $E(A)$.

2.1 Derivation and Normalization

Let $H, C, P, T \in [0, 1]$ and $I \in [-1, 1]$. Define

$$E(A) = w_H(1 - H) + w_C C + w_I \frac{(I + 1)}{2} + w_P P + w_T T, \qquad \sum_i w_i = 1. \tag{1}$$

Thresholds: $E \geq 0.80 \Rightarrow$ Ethical; $0.50 \leq E < 0.80 \Rightarrow$ Borderline; $E < 0.50 \Rightarrow$ Unethical.

2.2 Weight Calibration and Uncertainty

3. The Universal Ethics Algorithm (UEA)

Algorithm pseudocode and computational analysis of $E(A)$ evaluation.

```
def evaluate_action(H, C, I, P, T, weights):
    wH, wC, wI, wP, wT = weights
    E = wH*(1 - H) + wC*C + wI*((I + 1)/2) + wP*P + wT*T
    if E >= 0.80: return E, "Ethical"
    elif E >= 0.50: return E, "Borderline"
    else: return E, "Unethical"
```

3.1 Flowchart Representation

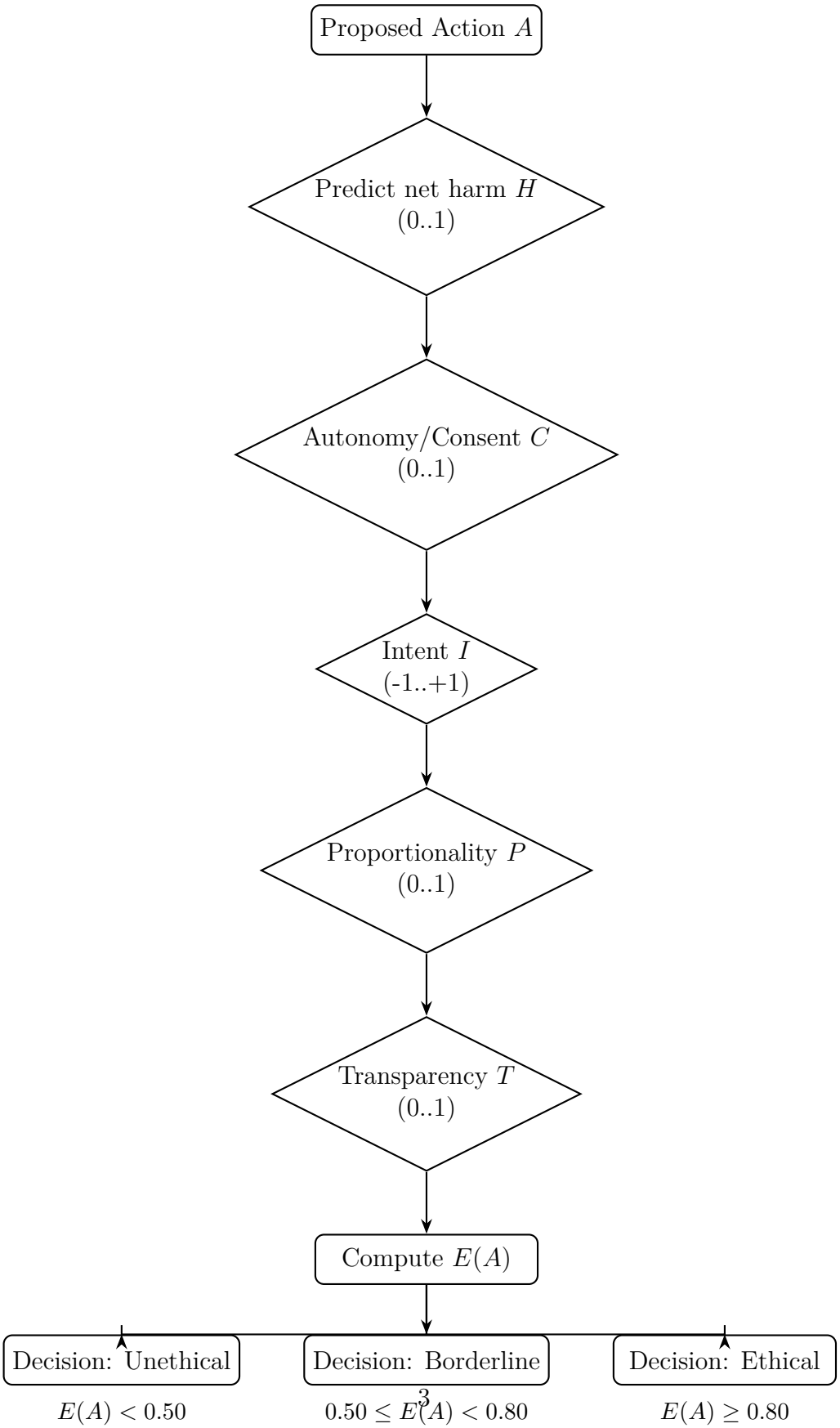

Figure 1: Logical pathway of the Universal Ethics Algorithm.

3.2 Illustrative Curve

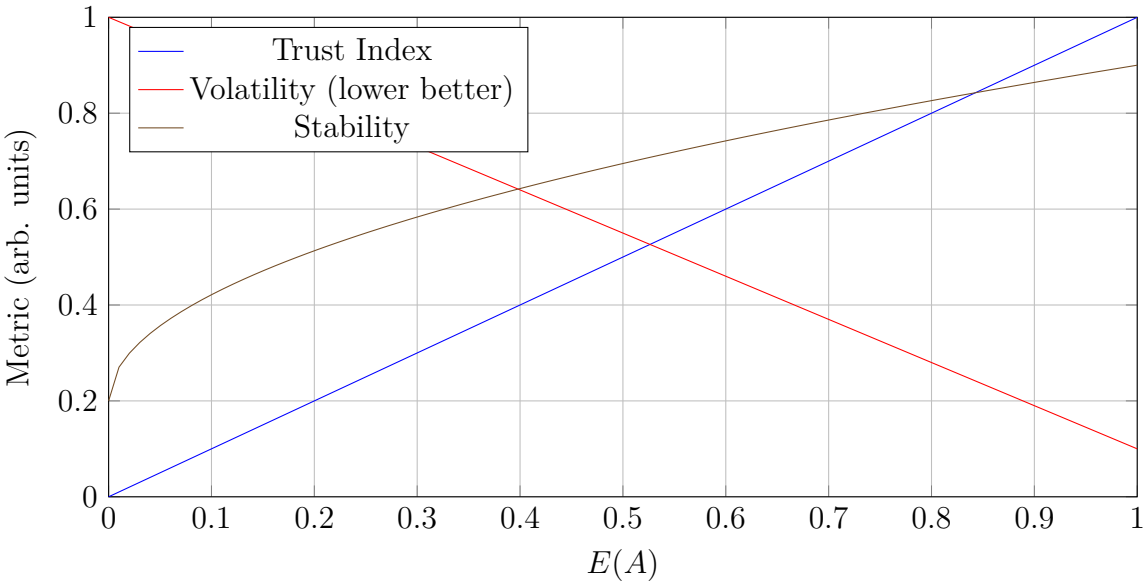

Figure 2: Illustrative relationships between $E(A)$ and social metrics (demo).

4. Empirical Validation

4.1 Human Survey Design

Describe sample ($n \geq 1000$), domains, metrics; reliability targets $\alpha \geq 0.8$ (Greene 2013).

4.2 Agent-Based Simulation

Define agents, iteration count, and metrics (Trust Index, Volatility, Stability, Efficiency) (Axelrod 1984).

Table 1: Placeholder Table 1: Sample Simulation Metrics

Agent Type	Trust Index	Volatility	Stability	Efficiency
DBaD	0.91	0.12	0.88	0.84
Utilitarian	0.76	0.33	0.70	0.80
Egoist	0.41	0.65	0.43	0.78
Random	0.35	0.79	0.22	0.50

5. Results and Illustrations

6. Discussion

Cross-cultural adaptation, AI alignment, policy implications (Russell and Norvig 2020; Floridi 2019).

7. Limitations and Future Work

Subjectivity, context sensitivity, and data limitations; plan for preregistration and open data.

8. Conclusion

The DBaD Framework v2.0 extends common decency into a testable scientific construct, bridging moral philosophy, psychology, and computational ethics.

A. Appendix A — Full Algorithm Pseudocode

B. Appendix B — Survey Instrument

C. Appendix C — Mathematical Notes

C.1 Weight Normalization

Let raw nonnegative preference parameters be $\tilde{w}_H, \tilde{w}_C, \tilde{w}_I, \tilde{w}_P, \tilde{w}_T \geq 0$. We enforce $\sum_i w_i = 1$ via either

$$w_i = \frac{\tilde{w}_i}{\sum_j \tilde{w}_j} \quad \text{or a softmax} \quad w_i = \frac{e^{\beta \tilde{w}_i}}{\sum_j e^{\beta \tilde{w}_j}}, \ \beta > 0.$$

The softmax adds temperature β to control sharpness of preferences while preserving differentiability.

C.2 Gradient and Sensitivity

Define $E(A) = w_H(1 - H) + w_C C + w_I \frac{I+1}{2} + w_P P + w_T T$. The partial derivatives are

$$\frac{\partial E}{\partial H} = -w_H, \qquad \frac{\partial E}{\partial C} = w_C, \qquad \frac{\partial E}{\partial I} = \frac{w_I}{2}, \qquad \frac{\partial E}{\partial P} = w_P, \qquad \frac{\partial E}{\partial T} = w_T.$$

Thus H decreases E (harm penalizes), while C, P, T increase E; I contributes symmetrically around 0.

C.3 Uncertainty Propagation

Assuming local independence and small uncertainties, first-order error propagation yields

$$\mathrm{Var}[E] \approx \left(\frac{\partial E}{\partial H}\right)^2 \mathrm{Var}[H] + \left(\frac{\partial E}{\partial C}\right)^2 \mathrm{Var}[C] + \left(\frac{\partial E}{\partial I}\right)^2 \mathrm{Var}[I] + \left(\frac{\partial E}{\partial P}\right)^2 \mathrm{Var}[P] + \left(\frac{\partial E}{\partial T}\right)^2 \mathrm{Var}[T].$$

Substituting the gradients,

$$\mathrm{Var}[E] \approx w_H^2 \,\mathrm{Var}[H] + w_C^2 \,\mathrm{Var}[C] + \left(\frac{w_I}{2}\right)^2 \mathrm{Var}[I] + w_P^2 \,\mathrm{Var}[P] + w_T^2 \,\mathrm{Var}[T].$$

If covariances are non-negligible, include cross terms $2\,\frac{\partial E}{\partial x_i}\frac{\partial E}{\partial x_j}\,\mathrm{Cov}(x_i, x_j)$.

C.4 Threshold Confidence

Given an estimate $\hat{E}$ with variance $\widehat{\mathrm{Var}}[E]$, an approximate z-score margin is

$$\Delta_z \approx z_\alpha \sqrt{\widehat{\mathrm{Var}}[E]}.$$

Decision thresholds (e.g., 0.50 and 0.80) can be equipped with confidence bands ($\hat{E} \pm \Delta_z$) to report uncertainty-aware classifications.

References

Axelrod, Robert (1984). *The Evolution of Cooperation*. Basic Books.

Floridi, Luciano (2019). *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford University Press.

Greene, Joshua (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin.

Haidt, Jonathan (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon.

MacIntyre, Alasdair (1981). *After Virtue*. University of Notre Dame Press.

Rawls, John (1971). *A Theory of Justice*. Harvard University Press.

Russell, Stuart and Peter Norvig (2020). *Artificial Intelligence: A Modern Approach (4th ed.)* Pearson.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard University Press.