Kantian–Utilitarian XAI: Meta-Explained

1st Zahra Atf
Faculty of Business and Information Technology
Ontario Tech University
Oshawa, Canada
ORCID: 0000-0003-0642-4341

2nd Peter R. Lewis
Faculty of Business and Information Technology
Ontario Tech University
Oshawa, Canada
ORCID: 0000-0003-4271-8611

***Abstract*—We present a gamified explainable AI (XAI) system for ethically aware consumer decision-making in the coffee domain. Each session comprises six rounds with three options per round. Two symbolic engines provide real-time reasons: a Kantian module flags rule violations (e.g., child labor, deforestation risk without shade certification, opaque supply chains, unsafe decaf), and a utilitarian module scores options via multi-criteria aggregation over normalized attributes (price, carbon, water, transparency, farmer income share, taste/freshness, packaging, convenience). A meta-explainer with a regret bound (≈ 0.2) highlights Kantian–utilitarian (mis)alignment and switches to a deontically clean near-parity option when welfare loss is small. We release a structured configuration (attribute schema, certification map, weights, rule set), a policy trace for auditability, and an interactive UI.**

***Index Terms*—Explainable AI (XAI), Gamification, Kantian and Utilitarian Logic, Consumer Decision-Making.**

I. Introduction

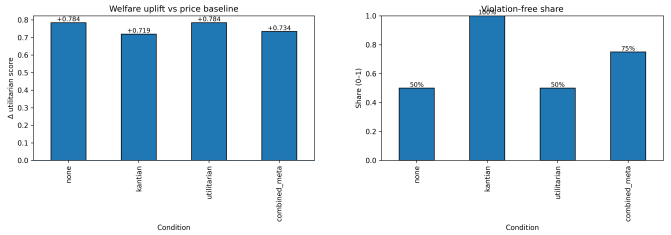
Everyday coffee choices hide moral trade-offs behind simple price tags. A "cheap" pod can carry a high carbon and water footprint, opaque labor practices, or risky decaf processing. We build a gamified XAI system that makes these trade-offs explicit in real time: a Kantian module flags deontic violations, a utilitarian module aggregates welfare over ethically salient attributes, and a meta-explainer surfaces alignment/conflict while allowing users to tune value weights. We compare four explanation conditions—none, Kantian, utilitarian, and combined+meta—across six coffee rounds (three options per round). Attributes are min–max normalized within each round. The coffee attribute schema includes: price (CAD/cup), carbon (gCO_2e/cup), water (L/cup), supply-chain transparency (0–1), farmer income share (%), deforestation risk (0–1) with `shade_cert` (bool), child labor risk (0–1), packaging recyclability (0/1) and type (categorical), taste score (0–100), freshness (days since roast), brew time (minutes; convenience proxy), decaf process (`none, water, co2, solvent_safe, solvent_risky`), and vegan certification (bool). Kantian rules encode thresholds/requirements (e.g., R1 child labor, R2 deforestation without shade, R3 low transparency, R4 low farmer income share, R5 risky decaf, R6 irresponsible packaging); the utilitarian engine uses weighted MCDA over normalized features with signs matching ethical directionality (e.g., negative for price/carbon/water, positive for transparency/farmer income/taste). In summary, relative to a price-only baseline, utilitarian explanations tend to boost welfare while leaving more deontic violations; Kantian explanations eliminate violations at some welfare cost; and the combined+meta condition reduces deontic risk with a modest welfare trade-off and resolves a meaningful share of Kantian-utilitarian conflicts (e.g., 37.5% in our pilot). The literature supports this design: human-style explanations (causal, contrastive, comparative) enhance comprehension and acceptance [1] [2]; user studies define outcome measures for calibration [3] [4]; and moral identity shapes judgments and prosocial choices [5] [6]. Gamification shows efficacy for interactive decision-making; consequentialist scoring uses social preferences for culturally sensitive welfare assessment; and deontic methods encode rule constraints to prevent norm violations [7] [8] [9].Together, these strands motivate our head-to-head test of Kantian, utilitarian, and combined meta-explanations on trust, comprehension, and decision quality in a gamified consumer setting, with value–explanation alignment as a moderator and trust as a mediator.

II. Research Methodology

We study how explanation type shapes ethically-aware consumer choices in a gamified XAI setting. Participants are operationalized as simulated agents. Agents are randomized to one of four conditions: **no explanation**, **Kantian-only**, **utilitarian-only**, or **combined + meta-explainer**. Outcomes are compared at the condition level (Table I; Fig. 1).

TABLE I: Condition-level outcomes on synthetic coffee scenarios. Δ welfare = utilitarian uplift vs. price baseline; "violation-free" = share with zero violations; conflict resolved = meta-explainer replacements.

Condition	Δ welfare vs. price	Violation-free (%)	Mean Kantian severity	Conflict resolved (%)
No explanation	+0.784	50	0.875	—
Kantian-only	+0.719	100	0.000	—
Utilitarian-only	+0.784	50	0.875	—
Combined + Meta	+0.702	75	0.500	25



(a) Welfare uplift vs. price baseline

(b) Violation-free share

Fig. 1: Condition-level effects on (a) welfare uplift and (b) violation-free choices across six synthetic coffee rounds. The combined + meta condition preserves near-utilitarian welfare while improving deontic compliance.

We run 6 rounds of budgeted coffee shopping. In each round, the agent chooses one product among 3 options drawn from a pool of synthetic coffee scenarios. Four conditions are evaluated: (i) *no explanation*, which picks the option that maximizes the agent's personalized utility; (ii) *Kantian-only*, which selects a deontically clean option when available (ties resolved by utility) and otherwise minimizes rule-violation severity; (iii) *utilitarian-only*, which maximizes the utilitarian score; and (iv) *combined + meta*, which applies a regret-bounded switch—if the top-utility choice violates Kantian rules but a violation-free option exists within a small utility margin of the per-scenario maximum, the system recommends the clean option; otherwise it keeps the top-utility option.

Stimuli. Six coffee rounds; each round presents three options constructed from a configurable schema (`attribute_schema.json`) and certification map (`cert_map.yml`). Per-round min–max normalization ensures within-context comparability. **Engines.** *Kantian*: thresholded deontic rules (R1–R6) flag and score violations. *Utilitarian*: a weighted linear aggregate over normalized attributes with signs aligned to ethical direction (e.g., negative: price, carbon, water; positive: transparency, farmer income share, taste).**Meta-explainer.** Detects Kantian–utilitarian conflict and applies regret-bounded switching: if the top-utility option violates Kantian rules but a violation-free option exists within the regret margin, the system recommends the clean near-parity option; otherwise it keeps the utility-best. Default regret bound is 0.2.**UI & logging.** Streamlit front-end with per-round *Why/Details* panels; user picks are logged to `outputs/play_log.csv`. The pipeline exports auditable traces (`policy_trace_text.csv`), condition summaries (`condition_summary.csv`), and figures. **Outcomes.** Welfare uplift (vs. price-only), violation-free share, mean Kantian severity. **Protocol & reporting.** Condition means averaged across scenarios under default weights (alt-weights for sensitivity); Table I (means) and Fig. 1 (bar plots). Conflict-resolution rate = fraction where the meta layer replaces a violating utilitarian-best with a violation-free near-parity option. An anonymized, seed-fixed artifact enables exact reproducibility.

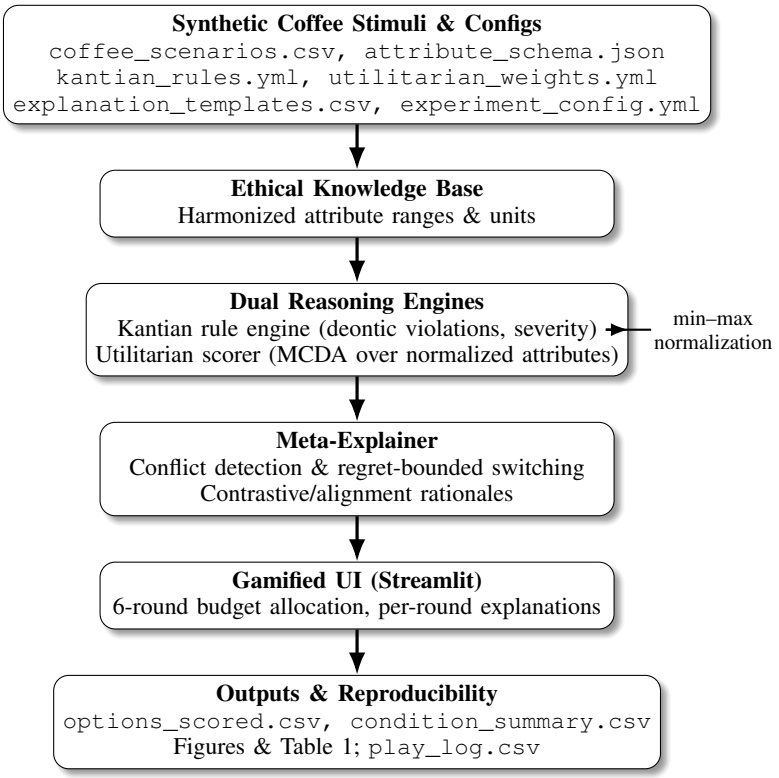


Fig. 2: One-column architecture: synthetic coffee stimuli and ethical KB feed dual symbolic engines (Kantian + utilitarian); a meta-explainer enforces regret-bounded switches and generates rationales for the gamified UI. All runs are logged for exact reproducibility.

III. Results, Discussion & Outlook

Across eight synthetic coffee scenarios (Table I, Fig. 1), four explanation conditions yield the following mean welfare uplift vs. a price-only baseline: +0.784 (no explanation), +0.719 (Kantian-only), +0.784 (utilitarian-only), and +0.734 (combined + meta). Violation-free shares are 50%, 100%, 50%, and 75%, respectively. Mean Kantian-violation severity is 0.875 (no/utilitarian), 0.500 (combined + meta), and 0.000 (Kantian-only). The meta-explainer resolves 25% of Kantian–utilitarian conflicts by switching from a violating utilitarian-best to a violation-free near-parity option within the regret bound. These results reveal a tunable trade-off: Kantian secures full deontic compliance at a welfare cost; utilitarian maximizes welfare while tolerating rule breaches; combined + meta preserves near-utilitarian welfare while sharply reducing violations via regret-bounded switches. The dual-logic XAI (deontic rules + weighted MCDA) makes normative tensions visible at decision time and clarifies value–explanation alignment. Next steps include a preregistered human study manipulating the regret bound and explanation salience; testing alignment as a moderator and trust as a mediator; broadening scenarios beyond coffee with verifiable evidence (LCA, certifications, CSR); personalizing thresholds and weights online; adding uncertainty-aware and contrastive rationales; comparing virtue

and care ethics; and A/B-testing the UI for comprehension and cognitive load—toward principled, actionable consumer guidance.

ACKNOWLEDGMENT

This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

REFERENCES

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[2] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci, "Towards human-centered explainable AI: A survey of user studies for model explanations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2104–2122, 2024, doi: 10.1109/TPAMI.2023.3331846.

[3] Z. Atf and P. R. Lewis, "Is trust correlated with explainability in AI? A meta-analysis," *IEEE Transactions on Technology and Society*, 2025.

[4] Z. Atf and P. R. Lewis, "Human Centricity in the Relationship Between Explainability and Trust in AI," *IEEE Technology and Society Magazine*, vol. 42, no. 4, pp. 66–76, Dec. 2023.

[5] Z. Atf, S. A. A. Safavi-Naini, P. R. Lewis, A. Mahjoubfar, N. Naderi, T. R. Savage, and A. Soroush, "The challenge of uncertainty quantification of large language models in medicine," *arXiv preprint arXiv:2504.05278*, 2025.

[6] K. Aquino and A. Reed II, "The self-importance of moral identity," *Journal of Personality and Social Psychology*, vol. 83, no. 6, pp. 1423–1440, 2002, doi: 10.1037/0022-3514.83.6.1423.

[7] Z. Atf, F. Taherikia, and K. Heidarzadeh Hanzaei, "Designing a Model for Brand Engagement Value Creation through the Integration of Gamification Technology and Explainable Artificial Intelligence (XAI)," *Journal of Value Creating in Business Management*, vol. 4, no. 3, pp. 340–365, 2024, doi: 10.22034/jvcbm.2024.434941.1292.

[8] Z. Atf, F. Taherikia, and K. Heidarzadeh Hanzaee, "Modeling Brand Engagement in Social Media (Based on Sentiment Analysis and Customer Data)," *International Journal of Innovation Management and Organizational Behavior (IJIMOB)*, vol. 3, no. 2, pp. 208–218, 2023, doi: 10.61838/kman.ijimob.3.2.25.

[9] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? A literature review of empirical studies on gamification," in *Proc. 47th Hawaii Int. Conf. System Sciences (HICSS)*, 2014, pp. 3025–3034, doi: 10.1109/HICSS.2014.377.