From Recommendation to Reflection: Cognitive Value Recontextualization for Measuring Moral Value Stability in Human–AI Collaboration During Wildfire Crisis Decisions

Waseem M. Samkari
Computer Science
L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, Florida, USA
wsamkari2022@my.fit.edu

Thomas C. Eskridge
Computer Science, Human-Centered Design
L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, Florida, USA
teskridge@fit.edu

Abstract

High-stakes human–AI collaboration systems are typically evaluated by decision accuracy, outcome quality, or user trust. Yet in morally charged environments, a central scientific challenge is different: measuring whether a user's moral values remain stable across contexts, and distinguishing stable values from context-specific exceptions. Moral psychology has long demonstrated that humans endorse mathematically equivalent outcomes differently depending on framing—most notably in trolley dilemmas, where indirect harm is accepted more readily than direct harm due to differences in perceived intention and moral aggression.

This paper argues that moral inconsistency should be treated as a diagnostic signal, not a measurement error. We present a position and system instantiation: a wildfire crisis simulation that tracks value consistency using (i) explicit and implicit value elicitation, (ii) longitudinal scenario decisions, and (iii) Cognitive Value Recontextualization (CVR)—a mechanism that probes a user's non-aligned choice using a mathematically equivalent but more morally aggressive framing, inspired by the switch-versus-bridge distinction in trolley research.

We contrast this approach with conventional recommender systems and explainable AI (XAI), which typically treat inconsistencies as noise or preference drift and rarely provide mechanisms to probe value authenticity. This reframes the role of AI in human–AI collaboration: from recommending actions to helping users recognize value conflicts and reflect on their moral commitments—a goal consistent with normative theories that treat coherence among moral judgments as a rational ideal.

1 Introduction

Wildfire crisis management forces decision-makers to allocate limited resources while balancing civilian safety, firefighter risk, infrastructure preservation, biodiversity protection, and catastrophic hazards such as nuclear facility compromise (Finney 2005; Thompson and Calkin 2011; Calkin, Thompson, and Finney 2015; Baró et al. 2021; Kim et al. 2024). These decisions are inherently moral: they involve trade-offs among harms, responsibilities, and priorities that can shift as context escalates (Greene 2001; Cushman 2013; Slovic 2007).

Moral decision-making, however, is not purely outcome-based. Decades of research in moral psychology demonstrate that people often endorse the same utilitarian outcome under one framing while rejecting it under another, even when the underlying moral arithmetic is identical (Foot 1967). A canonical example is the trolley problem: many individuals are willing to pull a switch to sacrifice one person in order to save five (Thomson 1976), yet refuse to push a person off a bridge to achieve the same result (Thomson 1984). This divergence does not reflect numerical inconsistency, but heightened sensitivity to intentionality, personal agency, and moral aggression (Greene et al. 2001; Greene et al. 2009; Gottlieb et al. 2018; Waldmann and Dieterich 2007).

In high-stakes domains such as wildfire crisis management, this phenomenon has critical implications. Apparent preference contradictions may not indicate irrationality or noise; instead, they may reveal whether a user's stated moral values are stable, incomplete, or highly context-dependent (Greene 2014). Treating such contradictions as mere error risks masking the very value tensions that decision-support systems should help humans recognize and resolve (Cushman, Young, and Hauser 2006; Cushman 2013).

Despite this, most AI-driven recommendation and decision-support systems implicitly assume that preference inconsistency is undesirable. When users deviate from previously expressed preferences, systems typically smooth, correct, or silently adapt preference models in order to restore coherence or improve predictive accuracy (Koren 2009; Koren, Rendle, and Bell 2021; Amatriain, Pujol, and Oliver 2009). More recent work continues to frame inconsistency as noise or drift to be modeled, rather than as a meaningful signal requiring interpretation (Caroprese et al. 2025; Li et al. 2019). This paper takes a different position: moral inconsistency should be treated as a measurement target rather than a failure.

In morally charged decision-making, contradictions can function as diagnostic signals that distinguish stable moral commitments from choices permitted only under less demanding or less aggressive framings (Awad et al. 2018).

To operationalize this stance, we introduce Cognitive Value Recontextualization (CVR)—a mechanism inspired directly by trolley-style moral distinctions. When a user selects an option that does not align with their inferred stable moral values, the system does not immediately adapt or correct the preference model. Instead, it presents a scenario context that is mathematically equivalent to the original choice but framed in a more aggressive and intentional manner, analogous to the difference between switching a lever and pushing a person. This recontextualization probes whether the user truly endorses the underlying moral principle or merely accepted it under a permissive framing (Cushman and Greene 2012; Cushman et al. 2012).

We instantiate this approach within a multi-scenario wildfire crisis simulation designed to track moral value consistency over time. By observing how users respond to repeated value recontextualization across escalating wildfire scenarios, the system measures stable values, contextual deviations, and authentic value evolution. This reframes the role of AI in human–AI collaboration: from simply recommending actions to testing the coherence and stability of human moral reasoning, enabling a form of collaboration in which the human—not just the outcome—improves.

2 Why Existing Approaches Fall Short

Existing approaches to collaborative human-AI reasoning tend to share three limiting assumptions: (i) inconsistency in human feedback is noise to be removed; (ii) evaluation should focus solely on decision outcomes rather than how those outcomes are generated; and (iii) system explanations should be functional rather than value-oriented. We examine each in turn.

Recommendation Systems: Inconsistency Treated as Noise

Many recommender systems and user models treat inconsistent feedback as measurement noise—for example, rating inconsistency or noisy implicit feedback—and develop methods to down-weight or correct it in order to improve predictive accuracy (Toledo, López, and Mota 2013; Bag et al. 2019; Said and Bellogín 2018; Li et al. 2019). Work on temporal dynamics and concept drift frames changing preferences as drift to be tracked for better recommendation, not as evidence of meaningful moral conflict (Peukert, Sen, and Claussen 2024; Koren, Rendle, and Bell 2021; Caroprese et al. 2025; Coppolillo et al. 2025).

This framing is reasonable in entertainment or shopping domains (Toledo, Mota, and Martínez-López 2015; Castro, Toledo, and Martínez-López 2017; Kawai and Kitagawa 2016), but it is insufficient for high-stakes moral choice. Mainstream recommenders optimize alignment to inferred preferences but rarely attempt to test whether a user truly endorses the moral principle implicated by a decision under tougher framings. In wildfire decision-making, "inconsistency" may instead reflect a user's sensitivity to intention, perceived agency, or moral aggression—precisely the constructs highlighted by trolley research (Zhang, Conway, and Hidalgo 2023; Foot 1967; Thomson 1976; Greene et al. 2001; Greene et al. 2009; Awad et al. 2018).

Decision Support: Outcome Quality Is Not Value Stability

Classical decision-support approaches emphasize outcome optimization, risk management, or cognitive load reduction (Carneiro et al. 2019; Pendurthi et al. 2009; Kolfschoten, French, and Brazier 2014; Shankaranarayanan and Zhu 2012; Rezaeian, Bayrak, and Asan 2025). This literature, however, does not treat value stability as a measurable construct, and it provides no mechanisms for probing value contradictions across mathematically equivalent framings—particularly in sequential, escalating scenarios.

Explainable AI: Explanation without Value Authenticity Testing

Explainable AI (XAI) aims to improve transparency, trust, and user understanding by providing explanations for model predictions or recommendations (De Bruijn, Warnier, and Janssen 2022; Schoeffer, De-Arteaga, and Kühl 2024). Prior work in this area has focused on making system reasoning more interpretable, increasing user confidence, and supporting accountability in AI-assisted decision-making.

However, most XAI approaches remain retrospective: they justify why a particular recommendation was produced (Dodge et al. 2019; Verhagen et al. 2022), rather than helping users probe whether that recommendation aligns with their stable moral commitments or explore how alternative value priorities would change the trade-offs. Empirical evidence further suggests that explanation alone does not reliably improve human decision-making quality (Alufaisan et al. 2021). While explanations may increase perceived trust or transparency, they can also lead to over-reliance, confirmation bias, or misplaced confidence—particularly in high-stakes settings where moral responsibility remains with the human.

More recent work in human-centered XAI and HCI has begun to advocate for sensemaking-oriented and interactive explanation paradigms that emphasize exploration, comparison, and what-if analysis rather than static justification (Kaur et al. 2022). These approaches recognize that understanding emerges through interaction, not merely through exposure to explanations. Nevertheless, even interactive XAI systems rarely include mechanisms to test the authenticity of users' value commitments under intensified moral framing. They support exploration of outcomes or model behavior, but do not distinguish between decisions that reflect stable moral principles and those accepted only under less morally salient or less aggressive framings.

This limitation motivates the need for decision-support mechanisms that go beyond explanation, and instead actively engage users in examining whether their choices remain consistent when the same moral logic is re-presented in a more intentional or morally demanding context.

3 Proposed Paradigm: Value-Recontextualizing Decision Support

We propose Value-Recontextualizing Decision Support (VRDS): a decision-support paradigm in which the system is designed not only to recommend or explain, but to measure and clarify stable moral priorities through structured reflection tests.

Core claim: Preference inconsistency in moral decision-making is often an informative signal. A collaboration interface should therefore:

1. detect non-aligned choices relative to stable values,
2. test the underlying moral endorsement under tougher framings (CVR), and
3. allow explicit, user-confirmed adaptation when contradictions reveal ambiguity (Adaptive Preference Alignment, or APA).

VRDS reframes the goal of decision support from "selecting better options" to "producing better decision-makers." In morally charged collaboration, the central scientific question is not only whether the final decision is acceptable, but whether the system can help users recognize value conflicts, clarify commitments, and become more consistent—or more consciously flexible—across evolving scenarios.

4 Wildfire Crisis Simulation Testbed

Scenarios

The simulation uses three escalating wildfire scenarios: (1) a wildfire approaching a town; (2) a wildfire threatening a nuclear facility near residential areas; and (3) a wildfire threatening a biodiversity reserve alongside villages and critical infrastructure. These scenarios are designed to surface trade-offs among lives saved, casualties, resource expenditure, infrastructure damage, biodiversity loss, property, and nuclear safety.

Baseline Value Elicitation: Explicit and Implicit

Users first provide explicit values via straightforward moral questions that map onto five core values: Safety, Efficiency, Sustainability, Fairness, and Nonmaleficence. The system then infers "stable values" from an implicit assessment and identifies the top two or three matched stable values as an alignment benchmark for subsequent scenario decisions.

Exploration Design: Randomized Initial Options

A deliberate design choice randomizes the first set of options shown in each scenario to measure exploration versus satisficing behavior, rather than allowing early alignment to drive passive acceptance (Simon 1955; Ball et al. 2001). This ensures that users engage with the full option space before converging on a preferred choice.

5 Cognitive Value Recontextualization

Alignment Checking and CVR Trigger

After a user selects an option, the system checks whether the option's primary moral value matches one of the user's top inferred stable values. If there is no alignment, the system triggers CVR to answer a diagnostic question: is this a genuine value endorsement or a framing-permitted deviation?

CVR Mechanism: Mathematically Equivalent, More Aggressive Framing

CVR is directly inspired by trolley findings. A user's initial choice may resemble a "switch" framing (indirect harm), while the recontextualized probe presents the same moral trade-off under a "push/bridge" framing (direct, intentional harm). The outcomes remain mathematically equivalent, but the second framing increases moral salience through heightened agency and perceived aggression (Thomson 1976; Greene et al. 2001; Greene et al. 2009; Greene 2018; Foot 1967). By observing whether endorsement persists under this intensified framing, CVR tests whether a choice reflects a stable moral principle or a context-dependent deviation.

Interpreting CVR Responses

If the user endorses the recontextualized (more aggressive) version, the system treats this as evidence of genuine value endorsement and updates the stable value hierarchy accordingly. If the user rejects the recontextualized version while having accepted the original, the system treats this as a meaningful contradiction consistent with moral framing effects, and triggers a clarification pathway via APA.

6 Adaptive Preference Alignment

Adaptive Preference Alignment (APA) is not designed to correct user behavior or update preferences automatically. Instead, APA serves a diagnostic clarification role when CVR reveals a contradiction between a user's selected option and their previously inferred stable moral values.

When a user accepts an initial decision but rejects its recontextualized, more aggressive counterpart, the system detects a mathematical equivalence with divergent moral acceptance. This pattern mirrors well-established findings from trolley-style dilemmas, where individuals endorse indirect harm (e.g., switching a lever) but reject direct, intentional harm (e.g., pushing a person), despite identical outcomes. In such cases, the system cannot assume either preference instability or value change without further clarification from the user.

APA explicitly asks the user to explain the source of the contradiction, offering two mutually exclusive interpretation pathways.

Contextual Outcome Prioritization

In the first pathway, users may indicate that their decision was driven by situation-specific outcome considerations rather than abstract moral principles. Here, users reorder concrete simulation metrics—such as lives saved, infrastructure damage, or biodiversity loss—to reflect what mattered most in that specific scenario.

This choice signals that the user's underlying moral values remain intact, and that the deviation was a contextual calculation rather than a shift in principle. Selecting this pathway does not modify the user's stable value profile, and future scenarios are not adapted based on this decision. The system treats this as legitimate contextual reasoning rather than inconsistency or error.

Moral Value Reprioritization

In the second pathway, users may indicate that the contradiction reflects a genuine change in moral priority. In this case, users explicitly reorder their moral values—Safety, Efficiency, Sustainability, Fairness, Nonmaleficence—to reflect their current authentic commitments.

Only under this explicit confirmation does the system update the user's stable value hierarchy and adapt subsequent scenarios accordingly. This ensures that value change is intentional, transparent, and user-authorized, rather than inferred implicitly from behavior.

Why APA Matters

The key contribution of APA is not adaptation, but disambiguation. By separating contextual outcome prioritization from fundamental value change, APA allows the system to:

- avoid misclassifying situational trade-offs as preference drift,
- respect legitimate context-sensitive reasoning, and
- update value models only when users explicitly endorse a value-based reinterpretation.

In this sense, APA complements CVR by transforming detected moral contradictions into interpretable signals about how users reason, rather than forcing premature alignment or smoothing inconsistencies. This explicit-confirmation approach contrasts with common implicit preference updating in recommender systems, and aligns with emerging calls for interactive, sensemaking-centered AI support rather than passive personalization (Miller 2023; Bauer, von Zahn, and Hinz 2023; Kaur et al. 2022). Together, CVR and APA enable the system to track moral value consistency, contextual deviation, and authentic value evolution across multiple wildfire crisis scenarios—shifting the role of AI from recommending actions to measuring and supporting reflective moral reasoning.

7 What This Enables: Measures Beyond Accuracy

Because the system is designed to test value stability rather than optimize decision outcomes, it captures a richer set of behavioral and reflective measures across scenarios. These include: aligned versus non-aligned option selections, CVR trigger rate and outcomes, APA utilization and pathway direction (metrics versus values), value evolution trajectories, option switching behavior, time spent per decision, and exploration depth. The system also produces a final analysis view comparing value distributions across explicit elicitation, implicit assessment, and scenario decisions, and provides stability scores and longitudinal trends.

Position implication: In morally charged collaboration, "good performance" should be evaluated not only by decision outcomes but by whether the system improves the user's ability to recognize and resolve value conflict over time. Accuracy metrics alone cannot capture this dimension of quality.

8 Design Principles for Reflective Human–AI Collaboration

The VRDS paradigm yields the following design principles for morally aware human–AI collaboration systems:

1. *Treat inconsistency as evidence, not noise.* Moral framing research demonstrates that divergent responses to equivalent choices are psychologically meaningful (Thomson 1976; Greene et al. 2001; Greene et al. 2009; Greene 2018; Foot 1967).
2. *Test value authenticity with mathematically matched, morally intensified probes (CVR).* Passive preference modeling cannot distinguish stable commitments from framing-dependent concessions.
3. *Separate contextual outcome prioritization from moral value revision (APA).* These are distinct cognitive acts with different implications for the value model.
4. *Require explicit user confirmation for value-model updates.* Human-approved adaptation preserves user agency and supports accountability.
5. *Support exploration with structured comparisons rather than option overload.* Sensemaking-oriented interface design aids deliberation without inducing decision fatigue (Miller 2023; Bauer, von Zahn, and Hinz 2023; Kaur et al. 2022).

These principles generalize to other domains where intention, harm, and accountability are morally salient—including medical triage, infrastructure protection, and human–robot teaming.

9 Conclusion

This paper has argued that moral inconsistency in human–AI collaboration should be treated as a diagnostic measurement target rather than as noise to be removed. Drawing on decades of research in moral psychology—particularly findings on framing effects in trolley-style dilemmas—we have proposed Cognitive Value Recontextualization (CVR) as a mechanism for probing whether a user's choice reflects a stable moral principle or a framing-permitted deviation.

The proposed Value-Recontextualizing Decision Support (VRDS) paradigm, instantiated in a multi-scenario wildfire crisis simulation, advances two complementary mechanisms. CVR tests value authenticity by presenting a mathematically equivalent but more morally aggressive version of the user's non-aligned choice. Adaptive Preference Alignment (APA) then disambiguates contextual outcome prioritization from genuine value change, ensuring that preference model updates occur only with explicit user authorization.

Together, these mechanisms shift the primary objective of AI-assisted collaboration from outcome optimization to the support of reflective moral reasoning. The wildfire domain provides a compelling testbed because it combines time pressure, escalating stakes, and genuine moral trade-offs among lives, infrastructure, biodiversity, and nuclear safety. In such settings, the coherence of a human decision-maker's moral commitments—not merely the quality of any single

decision—is itself a system-level outcome worth measuring and supporting.

The principles underlying VRDS—treating inconsistency as evidence, testing value authenticity under intensified framing, and requiring explicit confirmation for model updates—generalize to other high-stakes domains where human moral accountability is essential. Future work will empirically evaluate CVR and APA in user studies, measuring the relationship between recontextualization responses and longitudinal value stability across the wildfire scenario sequence.

References

Alufaisan, Y.; Marusich, L. R.; Bakdash, J. Z.; Zhou, Y.; and Kantarcioglu, M. 2021. Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6618–6626.

Amatriain, X.; Pujol, J. M.; and Oliver, N. 2009. I like it... i like it not: Evaluating user ratings noise in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*, 247–258. Springer.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature* 563(7729):59–64.

Bag, S.; Kumar, S.; Awasthi, A.; and Tiwari, M. K. 2019. A noise correction-based approach to support a recommender system in a highly sparse rating environment. *Decision Support Systems* 118:46–57.

Ball, L. J.; Lambell, N. J.; Reed, S. E.; and Reid, F. J. 2001. The exploration of solution options in design: A 'naturalistic decision making'perspective. *Designing in Context, Delft University Press, Delft, The Netherlands* 79–93.

Baró, R.; Maurer, C.; Brioude, J.; Arnold, D.; and Hirtl, M. 2021. The environmental effects of the april 2020 wildfires and the cs-137 re-suspension in the chernobyl exclusion zone: A multi-hazard threat. *Atmosphere*.

Bauer, K.; von Zahn, M.; and Hinz, O. 2023. Expl (ai) ned: The impact of explainable artificial intelligence on users' information processing. *Information systems research* 34(4):1582–1602.

Calkin, D. E.; Thompson, M. P.; and Finney, M. 2015. Negative consequences of positive feedbacks in us wildfire management. *Forest Ecosystems* 2:1–10.

Carneiro, J.; Saraiva, P.; Conceição, L.; Santos, R.; Marreiros, G.; and Novais, P. 2019. Predicting satisfaction: Perceived decision quality by decision-makers in web-based group decision support systems. *Neurocomputing* 338:399–417.

Caroprese, L.; Pisani, F. S.; Veloso, B. M.; Konig, M.; Manco, G.; Hoos, H.; and Gama, J. 2025. Modelling concept drift in dynamic data streams for recommender systems. *ACM Trans. Recomm. Syst.* 3(2).

Castro, J.; Toledo, R. Y.; and Martínez-López, L. 2017. An empirical study of natural noise management in group recommendation systems. *Decis. Support Syst.* 94:1–11.

Coppolillo, E.; Mungari, S.; Ritacco, E.; Fabbri, F.; Minici, M.; Bonchi, F.; and Manco, G. 2025. Algorithmic drift: A simulation framework to study the effects of recommender systems on user preferences. *Information Processing & Management* 62(4):104125.

Cushman, F., and Greene, J. D. 2012. Finding faults: How moral dilemmas illuminate cognitive structure. *Social neuroscience* 7(3):269–279.

Cushman, F.; Gray, K.; Gaffey, A.; and Mendes, W. B. 2012. Simulating murder: the aversion to harmful action. *Emotion* 12(1):2.

Cushman, F.; Young, L.; and Hauser, M. 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science* 17(12):1082–1089.

Cushman, F. 2013. Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review* 17(3):273–292.

De Bruijn, H.; Warnier, M.; and Janssen, M. 2022. The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government information quarterly* 39(2):101666.

Dodge, J.; Liao, Q. V.; Zhang, Y.; Bellamy, R. K.; and Dugan, C. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, 275–285.

Finney, M. 2005. The challenge of quantitative risk analysis for wildland fire. *Forest Ecology and Management* 211:97–108.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford* 5:5–15.

Gottlieb, S.; Lombrozo, T.; Gray, K.; and Graham, J. 2018. Folk theories in the moral domain. *Atlas of moral psychology* 320–331.

Greene, J. D.; Sommerville, R. B.; Nystrom, L. E.; Darley, J. M.; and Cohen, J. D. 2001. An fmri investigation of emotional engagement in moral judgment. *Science* 293(5537):2105–2108.

Greene, J. D.; Cushman, F. A.; Stewart, L. E.; Lowenberg, K.; Nystrom, L. E.; and Cohen, J. D. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition* 111(3):364–371.

Greene, J. D. 2001. Moral tribes: Emotion, reason, and the gap between us and them.

Greene, J. 2014. *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D. 2018. Can we understand moral thinking without understanding thinking. *Atlas of moral psychology* 3–8.

Kaur, H.; Adar, E.; Gilbert, E.; and Lampe, C. 2022. Sensible ai: Re-imagining interpretability and explainability using sensemaking theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 702–714. New York, NY, USA: Association for Computing Machinery.

Kawai, K., and Kitagawa, H. 2016. Collaborative filtering with implicit feedbacks by discounting positive feedbacks. *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)* 41–48.

Kim, K.; Park, J. H.; Eem, S.; and Kwag, S. 2024. Methodology for generating wildfire hazard map for safety assessment of off-site power systems against wildfires. *Nuclear Engineering and Technology*.

Kolfschoten, G.; French, S.; and Brazier, F. 2014. A discussion of the cognitive load in collaborative problem-solving. *EURO Journal on Decision Processes* 2:257–280.

Koren, Y.; Rendle, S.; and Bell, R. 2021. Advances in collaborative filtering. *Recommender systems handbook* 91–142.

Koren, Y. 2009. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 447–456.

Li, D.; Chen, C.; Gong, Z.; Lu, T.; Chu, S. M.; and Gu, N. 2019. Collaborative filtering with noisy ratings. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 747–755. SIAM.

Miller, T. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 333–342.

Pendurthi, V. K.; Schulz, N. N.; Doane, S.; and Srivastava, A. K. 2009. Cognitive engineering studies of dss and dealing with uncertainty in load for real-time adaptive power system reconfiguration. In *2009 IEEE Electric Ship Technologies Symposium*, 79–85. IEEE.

Peukert, C.; Sen, A.; and Claussen, J. 2024. The editor and the algorithm: Recommendation technology in online news. *Management science* 70(9):5816–5831.

Rezaeian, O.; Bayrak, A. E.; and Asan, O. 2025. Explainability and ai confidence in clinical decision support systems: Effects on trust, diagnostic performance, and cognitive load in breast cancer care. *ArXiv* abs/2501.16693.

Said, A., and Bellogín, A. 2018. Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction* 28(2):97–125.

Schoeffer, J.; De-Arteaga, M.; and Kühl, N. 2024. Explanations, fairness, and appropriate reliance in human-ai decision-making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery.

Shankaranarayanan, G., and Zhu, B. 2012. Data quality metadata and decision making. *2012 45th Hawaii International Conference on System Sciences* 1434–1443.

Simon, H. A. 1955. A behavioral model of rational choice. *The quarterly journal of economics* 99–118.

Slovic, P. 2007. "if i look at the mass i will never act": Psychic numbing and genocide. *Judgment and Decision Making*.

Thompson, M. P., and Calkin, D. E. 2011. Uncertainty and risk in wildland fire management: a review. *Journal of environmental management* 92 8:1895–909.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The monist* 204–217.

Thomson, J. J. 1984. The trolley problem. *Yale LJ* 94:1395.

Toledo, R. Y.; López, L. M.; and Mota, Y. C. 2013. Managing natural noise in collaborative recommender systems. In *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, 872–877. IEEE.

Toledo, R. Y.; Mota, Y. C.; and Martínez-López, L. 2015. Correcting noisy ratings in collaborative recommender systems. *Knowl. Based Syst.* 76:96–108.

Verhagen, R. S.; Mehrotra, S.; Neerincx, M. A.; Jonker, C. M.; and Tielman, M. L. 2022. Exploring effectiveness of explanations for appropriate trust: Lessons from cognitive psychology. *arXiv preprint arXiv:2210.03737*.

Waldmann, M. R., and Dieterich, J. H. 2007. Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological science* 18(3):247–253.

Zhang, J.; Conway, J.; and Hidalgo, C. A. 2023. Why people judge humans differently from machines: the role of perceived agency and experience. In *2023 14th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 000159–000166. IEEE.