Annotator Risk Preference as a Catalyst for Systemic Bias in Multimodal AI

Shuofeng Hu HUSHUOFENG08@163.COM
Zhen He HEZHEN.CS@GMAIL.COM
Tongtong Kan TONGTONGKAN@HOTMAIL.COM
Xiaomin Ying YINGXMBIO@FOXMAIL.COM
Beijing Institute of Basic Medical Sciences, Beijing, China

Abstract

As artificial intelligence evolves toward multimodal cognition, systems are moving beyond unimodal dependencies to integrate visual, auditory, and linguistic dimensions, thereby simulating human perception of reality. However, this increased complexity not only enhances expressiveness but also opens more insidious channels for bias infiltration. Existing research largely focuses on the demographic attributes of annotators (e.g., race, gender) while overlooking critical variables within the dimension of decision psychology (Ferrara, 2024; Sap et al., 2022). Among these, risk preference acts as a core driver of individual decision-making, exerting a subtle anchoring effect during the multimodal annotation process. When annotators confront materials characterized by high ambiguity, fuzziness, or potential social sensitivity, their intrinsic risk tolerance directly dictates label polarity, the degree of neutralization, and sensitivity toward minority attributes.

1. Introduction

With the deployment of Large Multimodal Models (LMMs) (Achiam et al., 2023), AI has deeply integrated into social production. To ensure model safety and utility, Reinforcement Learning from Human Feedback (RLHF) has become the standard paradigm (Ouyang et al., 2022). However, the industry faces the paradox of "the more alignment, the more bias" (Casper et al., 2023). Traditional views attribute this to long-tail data distributions or lack of diversity in annotation teams. This paper argues that even with perfectly balanced datasets, bias persists if the risk decision logic of annotators facing ambiguity is ignored. In multimodal tasks, the semantic ambiguity of images far exceeds that of text (Aroyo and Welty, 2015). Consequently, every annotation choice represents a trade-off between the *expected utility of accuracy* and the *quality control risk of error*. Thus, annotator risk preference—the tendency to commit Type I errors (over-rejection/false positives) versus Type II errors (omission/hallucination) under uncertainty—becomes the "invisible hand" shaping the model's value system.

2. Multimodal Bias Driven by Annotator Risk Preference

Risk Aversion and Over-Defensive Bias. In safety auditing or sensitive content filtering tasks, platforms often employ high-pressure penalty mechanisms. This induces extreme risk aversion among annotators. When encountering images with cultural specificity or non-mainstream content, risk-averse annotators tend to follow the "better safe than sorry" principle, labeling them as "NSFW" (Not Safe For Work) or refusing to describe them.

Once this decision logic is learned, it transforms into associative discrimination or spurious correlations (Geirhos et al., 2020). The model learns to strongly bind specific visual features (e.g., darker skin tones, specific cultural attire) with the concept of "risk." The result is a form of "cold deletion" during generation or retrieval, effectively constituting the digital erasure of marginalized groups, as observed in large-scale multimodal datasets (Birhane et al., 2021).

Risk Seeking and Stereotypical Hallucination. In contrast, in Image Captioning or Visual Question Answering (VQA) tasks, some annotators exhibit risk-seeking tendencies, particularly when visual information is occluded or missing. Instead of labeling a feature as "unclear" when a face is blurry or an object is obstructed, risk-seekers utilize social stereotypes to fill in the blanks. For instance, a blurry figure in a nurse's uniform is labeled "female". This gambling-style annotation leads the model to ignore visual evidence during cross-modal alignment, instead relying on linguistic statistical regularities (Rohrbach et al., 2018). Bias is not only preserved but hallucinated—the model confidently generates details that do not exist in the image but align with stereotypes (Li et al., 2023).

The Consensus Trap and Neutrality Bias. Furthermore, the consensus mechanisms of crowd-sourcing platforms exacerbate these issues. To maximize earnings, annotators attempt to predict the majority vote (Geva et al., 2019). When facing ambiguous cases, risk-averse annotators tend to select "neutral" or "uncertain" tags to minimize the probability of being flagged as a "false positive" by quality control systems. This is termed "neutrality bias" (Voutsa et al., 2025). Here, neutrality is not an objective reflection of reality but a defensive decision. This renders the model extremely obtuse when processing real-world edge cases or subtle bias signals (such as micro-aggressions), as the training signals have been heavily masked by defensive neutralization.

3. Algorithmic Interventions

To address this, one approach involves introducing annotator characteristics (specifically the risk coefficient, α) into the training loop. By feeding α into the network alongside the input sample (text + image), effective inductive bias is introduced, enabling the model to discern the objectivity of the description.This aligns with recent work on modeling individual annotator perspectives rather than aggregating them into a single ground truth (Davani et al., 2022). Another method involves dynamic modality masking during the alignment phase. If the system detects an over-reliance on a modality heavily influenced by annotator bias (e.g., the visual channel), it can dynamically mask that modality, forcing the model to seek features from complementary, less biased modalities (e.g., text) (Cadene et al., 2019).

4. Position and Future Directions

Annotator risk preference is an underestimated phantom variable in the genesis of multimodal AI bias. Through cognitive anchoring, defensive neutralization, and loss aversion mechanisms, it fundamentally alters the substructure of labeled data, transforming objective uncertainty into systemic bias laden with subjective value judgments.

To rectify this, the technical community must transcend the data cleaning mindset and pivot toward rational mechanism design:

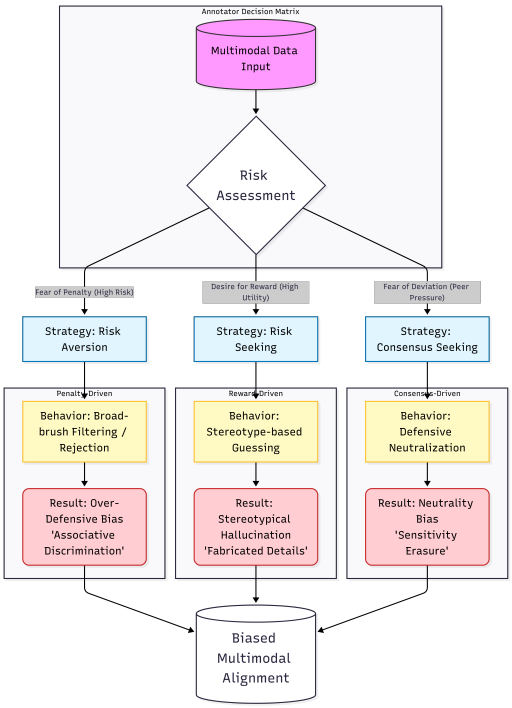

Figure 1: The propagation mechanism of multimodal bias driven by annotator risk preference.

- **Introduce Epistemic Uncertainty Labels**: Abandon forced binary choices in subjective tasks in favor of soft labeling paradigms that preserve the full distribution of uncertainty.
- **Decouple Reward Loops**: Separate the reward functions for safety and utility to prevent risk aversion from causing over-censorship or risk seeking from causing hallucination.
- **Psychological Profiling**: Establish dynamic psychological profiles for professional annotation teams, treating risk preference as essential metadata.
- **Adversarial Risk Training**: Introduce expert models with varying risk preferences during the RLHF stage to act as adversaries, forcing the primary model to dynamically adjust its confidence thresholds across different contexts.

Through these systemic interventions, we can not only enhance the precision of AI but also construct an algorithmic system of fairness that possesses *humanized prudence* while transcending *human bias*. This represents not merely a technical victory, but a profound reshaping of the ethical governance path for artificial intelligence.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*, 2019.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906, 2022.

Maria C Voutsa, Nicolas Tsapatsoulis, and Constantinos Djouvas. Biased by design? evaluating bias and behavioral diversity in llm annotation of real-world and synthetic hotel reviews. *AI*, 6(8):178, 2025.