The MISOMEM-Val Dataset for Identifying Human Values in Misogynistic Memes

Rakshitha Rao Ailneni, Sanda Harabagiu
Human Language Technology Research Institute, The University of Texas at Dallas
Richardson, TX, USA
{rxa220074, sanda}@utdallas.edu

Abstract

We present MISOMEM-Val, the first dataset that systematically annotates human values across Frames of Misogyny (FoMs) derived from misogynistic memes. Extending the Taxonomy of Misogyny, each frame is linked to the Human Value Hierarchy (HVH) with annotated support and ignore stances and accompanying rationales. In total, 1089 frames were annotated, presenting 3,051 support and 7,007 ignore stance value towards various FoMs. We introduce Hierarchical Value Annotation with Human Feedback (HVA-HF), an LLM-assisted annotation framework combining Chain-of-Thought prompting and self-consistency verification to ensure transparency and quality. The annotation analysis reveals systematic asymmetries: values of Conservation and Self-Enhancement are frequently supported, while Self-Transcendence is often ignored, thus highlighting how misogynistic memes distort core human values.

Disclaimer: This paper contains examples of misogynistic and hateful content that may be disturbing to some readers.

Keywords: Human values, Framing, Misogyny, Large Language Models

1. Introduction

Social scientists have long argued that people's basic values predict their attitudes and behavior (Rokeach, 1974; Ball-Rokeach et al., 1986). (Schwartz, 1992) sought to identify a comprehensive set of ten human values that are recognized in all societies, serving as guiding principles in the life of a person or group. In (Schwartz et al., 2012), the set of human values was refined and organized in a value taxonomy. This taxonomy was further enhanced in (Kiesel et al., 2022), organizing 54 human values across three different hierarchical levels. This taxonomy was used to automatically identify human values in textual arguments in the SemEval-2023 Task 4 (Kiesel et al., 2023).
However, human values do not guide only the textual argumentative communications. Human values also govern the generation of communications that are combining images with texts, as in the case of *m*emes. Memes have become a prevalent means of online communication (Joshi et al., 2023) because they are generally amusing and quick to consume. But memes can often be malicious, propagating hate. For example, Figure 1 illustrates a meme that propagates misogyny in the form of stereotyping. Misogyny, which is defined as hatred of, aversion to, or prejudice against women, is manifested in multiple forms, stereotyping being just one of them.
The discovery of human values elicited by misogynistic memes is important because it allows us to understand (a) what motivates this form of hate and (b) how misogyny is hurting women by disregarding their own human values. For the meme illustrated in Figure 1, misogyny is motivated by *conservatism* values that invoke *security* concerns, framing women's act of driving as a threat to societal order and stability. While the meme creator is observing these conservative values, it purposely ignores *s*elf-transcendence values by not being

Figure 1: Meme showcasing misogyny and the Human Values it elicits.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4031–4047
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

committed to equality between genders (e.g., universalism). In addition to highlighting the motivations and harm provoked to women, the discovery of human values elicited by misogynistic memes has the potential to inform counter-hate arguments (Alyahya and Aldayel, 2024; Albanyan et al., 2023) that effectively fight and limit the spread of hateful messages as they become more persuasive.

Since there are no available datasets annotating the human values elicited by memes, we created the first such dataset, containing MISOgyny MEMes (MISOMEM) on which Human Values were annotated, generating the **MISOMEM-Val** dataset. In generating MISOMEM-Val, we took advantage of an existing meme benchmark used in the SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022). As it can be seen from Figure 1, it is not easy to infer the human values elicited by a meme where complex images and superimposed text are combined. First, one has to recognize the form of misogyny that is used in the meme: stereotyping, which is defined as the practice of assigning a fixed, conventional idea or set of characteristics to a woman. Then we need to be aware of how misogyny is framed by the meme, namely that women are bad drivers, since in a world with no men, traffic chaos would emerge, as seen in the lower image of the meme. Therefore, we found that instead of inferring directly from a meme the human values it elicits, it is much easier to discover these values from the way misogyny is framed by the meme. Fortunately, we had access both to the misogyny type and to the way in which misogyny was framed in each meme from the MAMI dataset through the data made available in (Ailneni and Harabagiu, 2025).

The work reported in (Ailneni and Harabagiu, 2025) introduced two concepts: (1) Misogyny Problems (MPs), representing various types of misogyny encountered in memes, and (2) Frames of Misogyny (FoMs). FoMs are a special form of framing. Framing is a concept central to communication sciences. The definition of framing provided in Entman (1993) notes that *"to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."* Based on this definition, the Misogyny Problems (MPs), e.g., stereotyping, shaming, or objectification, represent the aspects of misogyny that receive a causal interpretation in the articulation of FoMs. Figure 2 illustrates (a) the MP addressed by the meme showcased in Figure 1 as well as (b) the FoM that it evokes. The figure also shows the human values elicited by the FoM evoked by the meme.

In MISOMEM-Val, each meme is annotated

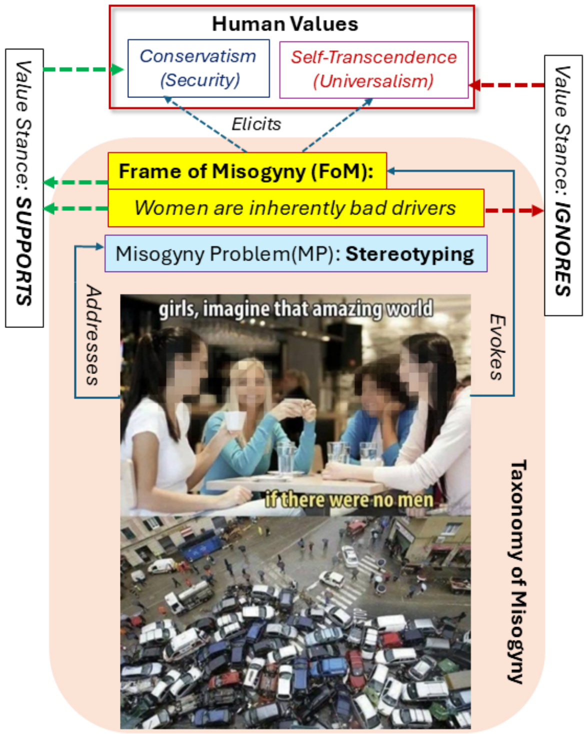

Figure 2: Misogynistic meme, the Misogyny Problem (MP) it addressed, and the Frame of Misogyny (FoM) it evokes in the Taxonomy of Misogyny (ToM), the Human Values (HMs) elicited by the FoM, and the stance values of the HMs towards the FoM.

with (1) information available from the Taxonomy of Misogyny (ToM) released by (Ailneni and Harabagiu, 2025), namely the MP it addressed and the FoM it evokes, and (2) all human values elicited by the FoM. From the ToM, we had access to 1089 FoMs *evoked* by 10K memes, in which 99 different MPs were *addressed*. Moreover, the ToM linked together all memes evoking the same FoM, as well as all FoMs that addressed the same MP. We note that, as shown in Figure 3, the structure of the ToM made the annotation of the human values very efficient, because the human values that were discovered to be elicited by each FoM_x were considered to be elicited by all memes $\{m_i\}$ that evoked FoM_x.

Recently, (Borenstein et al., 2025) has shown the importance of also considering the *stance* towards human values. Stance is defined by Biber and Finegan (1988) as the expression of an author's standpoint and judgment towards a given proposition. Moreover, the stance always has a subject and an object. The subject of stance can be the speaker in a conversation, the author of a social media post, or the meme author, in our case. The stance object, as reported in (Hardalov et al., 2021; Liu et al., 2023), can be a controversial entity, concept, idea, event, article headline, or claim. In (Weinzierl and Harabagiu, 2024b), the role of the object of the stance was highlighted. For MISOMEM-Val, we

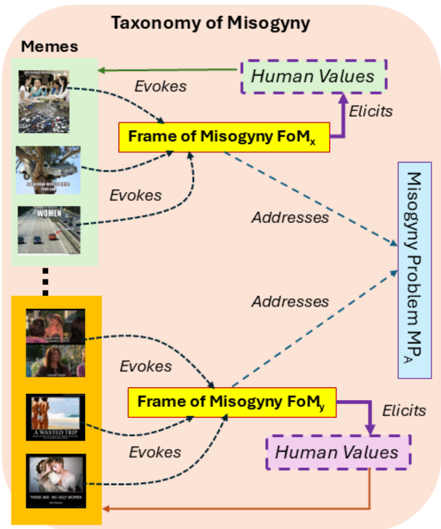

Figure 3: MISOMEM-Val annotation of human values on memes from the Taxonomy of Misogyny (ToM).

consider the object of the stance as the human value it elicited by a meme. Therefore, we measure the *stance* of a human value h_a towards the frame of misogyny FoM_i evoked by a meme m_j. Moreover, when judging the stance, we can attribute one of two possible values, namely *supports* and *ignores*. In Figure 2 we illustrate the values of the stance of each human value elicited by the FoM evoked in the illustrated meme.

In annotating the FoMs with human values, we adopt the Human Value Hierarchy (HVH) (Kiesel et al., 2022), illustrated in Figure 4. Each FoM is annotated across three levels of granularity: higher-order values (level 3), value categories (level 2), and specific values (level 1). In addition, we capture the *rationale* explaining (a) the human values annotated for each meme and (b) the stance value of the human value towards the FoM evoked by the meme. These rationales explain both the human values and the stance values for each meme. In this paper, we make the following contributions:
◁**1**▷ We introduce **MISOMEM-Val**, the first dataset in which human values are systematically annotated on memes. Each annotated human value is provided with (a) its stance value towards the Frame of Misogyny (FoM) evoked by the meme and (b) a rationale for the selection of the human value from the Human Value Hierarchy (HVH) and for the stance associated with the human value.
◁**2**▷ We propose a semi-automated, LLM-assisted methodology for annotating and curating human values from misogyny memes, namely the **Hierarchical Value Annotation with Human Feedback (HVA-HF)**. HVA-HF annotates human values and their stance value towards the FoM evoked by each meme. We make available the MISOMEM-Val dataset as well as prompts that we used on GitHub[1].
◁**3**▷ We present several methods for learning to automatically identify human values in memes, based on the annotations available from MISOMEM-Val.

2. Related Work & Background

2.1. The Human Value Hierarchy

To identify human values elicited by a Frame of Misogyny (FoM), we rely on the Human Value Hierarchy (HVH) introduced by (Kiesel et al., 2022), detailed in Figure 4. HVH integrates 54 human values derived from four major cross-cultural sources: the Schwartz Value Survey (Schwartz et al., 2012), Rokeach Value Survey (Rokeach, 1973), Life Values Inventory (Brown and Crace, 2002), and World Values Survey (Haerpfer et al., 2020). HVH comprises four levels: Level 1 lists 54 Human Values (e.g., be creative, have an exciting life) (Schwartz, 1994); Level 2 groups them into 20 Value Categories: 19 from Schwartz's taxonomy plus one added by (Kiesel et al., 2022); Level 3 defines four Higher-Order Values, shown in Figure 4; and Level 4 encodes base dichotomies, distinguishing personal vs. social focus (4a) and anxiety-based vs. anxiety-free motivation (4b). For our annotations, we have considered only levels 1, 2 and 3 of the HVH. Moreover, we took into account that the HVH contains multiple possible *value paths* of the form $HV_A \rightarrow HV_B \rightarrow HV_C$, where HV_A is a value from Level 3 of HVH, HV_B is a value from Level 2, and HV_C is a value from Level 1. Toi be noted that due to hierarchical relations in HVH we know that HV_C has HV_B as its parent and HV_B has HV_A as its parent. One example Value PAth (VP) is: *Self-Transcendence* $\rightarrow$ *humility* $\rightarrow$ *have life accepted as is*, as derived from the HMV illustrated in Figure 4.

2.2. Prior Work targeting human value annotations

Value systems have been applied in formal argumentation to model audience preferences through value-based argumentation schemes (van der Weide et al., 2010), defeasible logic programming (Teze et al., 2019), and value-based frameworks (Bench-Capon, 2003). Central to these approaches is the automatic identification of values in text. (Kiesel et al., 2022) introduced the Webis-ArgValues-22 dataset with 20 categories and 54 values, later extended by (Mirzakhmedova et al., 2024) to 9,324 arguments in Touché23-ValueEval. Beyond argumentation, values have been widely explored in NLP: (Ammanabrolu et al.,

[1] https://github.com/rak55/MISOMEM-Val

2022) guided conversational agents toward moral behavior, while (Qiu) proposed ValueNet, containing 21,000 SOCIAL-CHEM-101 (Forbes et al., 2020) situations annotated with ten value categories. (Borenstein et al., 2025) developed a transformer-based framework to predict the relevance and polarity of human values across nine million Reddit posts. However, automatic value identification has so far been limited to text-based content, with no prior work addressing human values in memes, particularly misogynistic ones.

2.3. A Taxonomy of Misogyny

(Ailneni and Harabagiu, 2025) present the first automatic method for articulating Frames of Misogyny (FoMs) from memes, without presupposing knowledge of all misogyny problems (MPs) represented in SMPs. Their framework simultaneously induces and organizes the discovered MPs into a Taxonomy of Misogyny (ToM). The ToM was generated on the dataset introduced by (Fersini et al., 2022), which consists of 11,000 memes collected from X/Twitter, Reddit, 9GAG, KnowYourMeme, and Imgur. While these memes were originally annotated with four high-level MPs—*Stereotyping, Shaming, Objectification, and Violence*—the method in (Ailneni and Harabagiu, 2025), using several few-shot Chain-of-Thought (CoT) prompting (Wei et al., 2022b), uncovers **1089** unique frames and **99** unique MPs. These MPs are further structured into 11 top-level hierarchies: *Objectification, Stereotyping, Violence, Shaming, Patriarchal Attitudes, Disrespect toward Women, Trivializing Serious Issues, Discrimination of Women, Dehumanization of Women, Exploitation of Women by Men, and Pseudoscience*. These hierarchies, along with the FoMs evoked by memes create the ToM, which we have also used in this work.

3. MISOMEM-Val Annotations

We believe that identifying the human values $\{HV_i\}$ elicited by a meme M_i can be performed by discovering the human values elicited by the Frame of Misogyny FoM_j evoked by M_i. Therefore, the visual and textual content of memes influences the annotation process only indirectly, since our annotations rely mainly on the FoM evoked by the meme, available from the ToM. We note that individual memes may evoke multiple FoMs, each providing its annotations of human values. Our proposed method, **Hierarchical Value Annotation with Human Feedback (HVA-HF)**, operates in two different phases:

Phase A concerns with the annotation of (a) entire *Value Paths* $\{VP_k\}$, available from the HVH, on each FoM_j evoked by one or several memes $\{M_i\}$; and (b) the *stance* of values from each VP_k towards FoM_j. Our annotations of human values consider the ontological commitments from HVH, providing annotations at three levels of abstraction. In addition, we use the observation that all values from the same VP_k share the same stance towards FoM_j, either supporting it or ignoring it. Appendix A illustrates examples of memes, FoMs they evoke, value paths annotated as well as their stance.

L3	L2	L1
Openness to change	Self-direction: thought	Be creative Be curious Have freedom of thought
	Self-direction: action	Be choosing own goals Be independent Have freedom of action Have privacy
	Stimulation	Have an exciting life Have a varied life Be daring
	Hedonism	Have pleasure
Self-enhancement	Achievement	Be ambitious Have success Be capable Be intellectual Be courageous
	Power: dominance	Have influence Have the right to command
	Power: resources	Have wealth
Conservation	Face	Have a social recognition Have a good reputation
	Security: personal	Have a sense of belonging Have a good health Have no debts Be neat and tidy Have a comfortable life
	Security: societal	Have a safe country Have a stable society
	Tradition	Be respecting traditions Be holding religious faith
	Conformity: rules	Be compliant Be self-disciplined Be behaving properly
	Conformity: interpersonal	Be polite Be honoring elders
Self-transcendence	Humility	Be humble Have life accepted as is
	Benevolence: caring	Be helpful Be honest Be forgiving Have the own family secured Be loving
	Benevolence: dependability	Be responsible Have loyalty towards friends
	Universalism: concern	Have equality Be just Have a world at peace
	Universalism: nature	Be protecting the environment Have harmony with nature Have a world of beauty
	Universalism: tolerance	Be broadminded Have the wisdom to accept others
	Universalism: objectivity	Be logical Have an objective view

Figure 4: The Human Value Hierarchy (Kiesel et al., 2022) encodes at Level 1 (L1): 54 Human Values (HVs); at Level 2 (L2): 20 Value Categories (VCs) and at Level 3 (L3): 4 Higher Order Values (HOVs).

Phase B focuses on the Verification and Editing of annotations.

3.1. Phase A: Annotation of Value Paths and their Stance

To produce annotations pairing Value Paths with their stance towards the FoM evoked by a meme, we took advantage of the demonstrated capacity of Large Language Models (LLMs) as effective zero-shot reasoners (Kojima et al., 2022) Specifically, we prompt **GPT-5-mini** using Chain-of-Thought (CoT) reasoning (Wei et al., 2022a). We selected GPT-5-mini as the backbone model because it provides sufficient reasoning quality for discovering Value Paths while remaining computationally efficient for large-scale annotation. For each FoM_x, the LLM is provided with (a) the Misogyny Problems $\{MP_i^x\}$ addressed by FoM_x along with their definitions, available from ToM (Ailneni and Harabagiu, 2025) and (b) the complete HVH and its definitions of values (Schwartz et al., 2012). The LLM is instructed to output (a) the *Value Paths (VPs)* and (b) the corresponding *stance* of the values from a VP towards FoM_x along with detailed rationales for the discoveries it makes.

Rather than instructing the LLM to identify only human values, encoded at Level 1 of HVH, we adopt a hierarchical, abstract-to-concrete discovery strategy (Yoshimura and Kashima, 2025; Lo et al., 2023; Budagam et al., 2025) for annotating the human values. The LLM is first prompted to identify the Higher-Order Values from Level 3 of the HVH that are elicited by some FoM_x. Once these Higher-Order Values $\{HOV_i\}$ have been generated, the LLM is subsequently guided to identify $\{VC_k\}$, the Value Categories (Level 2) and, ultimately, $\{HV_K\}$ the human values (Level 1). This enabled the generation of Value Paths $\{VP_y\}$ when considering the hierarchical relations encoded in HVH between (a) the elements of the $\{HOV_i\}$ set and the elements of the $\{VC_j\}$ set; and (b) the relations between the elements of the $\{VC_j\}$ set and those of the $\{HV_k\}$ set. Then, for each Value Path VP_y the LLM is prompted to generate the stance of VP_y towards the FoM_x. In this phase, the LLM also generates rationales that justify both the Value Paths and their their stance. Detailed system prompt is shown in Appendix B.

To further enhance the reliability of annotations, we incorporate self-consistency both as a decoding strategy and as a confidence signal for the LLM's final outputs (Wang et al., 2023). For each FoM_x, we generate ten independent reasoning samples, and compute the confidence in a Value Path VP_x^i and the value of its *stance* based on its frequency across these samples. Paths that appear in more than eight out of ten samples are considered high-confidence, while the remaining paths are flagged for human verification. This procedure substantially reduces annotation effort by automatically filtering robust model predictions while maintaining strict quality control for uncertain cases.

3.2. Phase B: Verify and Edit

In this phase, low-confidence human value paths and their stance are verified to ensure that the identified human values are indeed elicited by the corresponding FoM. The annotators review and revise both human value paths and the predicted stance values and the accompanying rationales, when necessary.

Annotator guidelines: We recruit three annotators with near-native English proficiency and a computer science background to verify the generated human values. We recruited annotators with a computer science (NLP) background because the task required understanding the hierarchical structure of the Human Value Hierarchy, interpreting LLM-generated outputs, and consistently verifying value paths and their stance. This technical background helped ensure annotation consistency and reliable verification of model-generated rationales. Initially, we used the annotation guidelines showcased in Figure 5, instructing the annotators to select the human values they believe were elicited by the FoM, along with the appropriate stance value towards the FoM. These guidelines enabled us to conduct a pilot study. In this study, annotators labeled value paths—along with their *support* or *ignore* stance values—for 50 randomly sampled FoMs and also provided rationales for each value path. While performing their annotations, the annotators do not have access to one another's annotations, ensuring that each annotation is produced independently.

At the end of the pilot, all annotators meet to discuss the cases where value paths and/or stances were not annotated in the same way and to compare rationales. Through this discussion, the annotators agree on three criteria that every rationale should satisfy: (1) *Alignment with the value definition*, meaning that the rationale for a particular human value from HVH should be consistent with the original value definition in (Schwartz et al., 2012); (2) *Applicability to the FoM evoked by the meme*, requiring the rationale to be specific and relevant to the given FoM rather than vague or generic; and (3) *Adherence to the stance*, requiring that the rationale explicitly justify whether the value is *supported* or *ignored* by the FoM. The edited rationale shown in Figure 6 exemplifies these criteria: it explicitly grounds the explanation in the FoM by highlighting women's right to choose relationships (criterion 2), maintains fidelity to the HVH definition of Self-direction (action) (criterion 1), and clearly frames the opposition as an *ignore* stance (criterion 3).

The inter-annotator agreement (IAA), measured

Objective:
Verify and refine each **low-confidence** *(Value, Stance)* path generated by the LLM. Begin by assessing whether the value path is relevant and applicable to the given **Frame of Misogyny (FoM)**. If the value path is determined to be applicable, proceed to verify and edit the corresponding **stance** and the **rationale** generated alongside the value path.
Low-confidence:
The confidence of a *(value, stance)* path is estimated using self-consistency decoding, wherein multiple generations are sampled from the LLM, and the confidence score is computed based on the frequency of the path's occurrence across these sampled generations. A higher frequency reflects greater confidence, and vice versa.
Annotation steps:
1. **Value Verification:**
   - Review the FoM and its associated misogyny problems (MPs).
   - Assess whether the low-confidence **value path** generated by the LLM is applicable to the FoM. Refer to the **Human Value Hierarchy (HVH)** for an accurate definition of the value.
2. **Stance Verification & Editing:**
   - If the value path is deemed applicable to the FoM (from Step 1), evaluate whether the stance generated alongside the value path is correct.
   - If the stance is incorrectly generated, edit the stance.
3. **Rationale Verification & Editing:**
   - Alignment with Value Definition: Ensure the rationale accurately reflects the value definition from the HVH.
   - Applicability to the FoM: Confirm that the rationale directly relates to the FoM and its MPs, avoiding vague or generic statements.
   - Adherence to Stance: Ensure the rationale clearly justifies the stance.

If the rationale fails to meet any of these criteria, revise the rationale so that all the three criteria are met.

Figure 5: Annotator guidelines.

as Krippendorff's alpha (Krippendorff, 2011), was 0.83 for annotating 50 FoMs with stance and 0.79 for value paths. To estimate inter-annotator agreement for rationales, which are free-form text, each annotator rated the rationales written by the other two annotators on a five-point Likert scale grounded in the three rationale criteria (alignment with the HVH value definition, applicability to the FoM, and adherence to the assigned stance). The scale was defined as follows: (1) the rationale is not relevant to the FoM and does not align with the value defi-

Frame of Misogyny: Women are single because they don't want good men.
Value: Self-direction (action) (Level 2)
Value Definition: Freedom to determine one's own actions

Generated Rationale: Opposes the value of people determining their own actions and pursuing personal goals.

Edited Rationale: The frame dismisses women's right to choose their relationships freely, opposing the value of people determining their own actions and pursuing personal goals.

Figure 6: Example showing a rationale edited by the annotators.

nition; (2) the rationale is relevant to the FoM but does not justify the selected value based on the HVH definition; (3) the rationale is relevant to both the FoM and the value definition but does not clearly justify the assigned stance; (4) the rationale is relevant to the FoM, aligned with the value definition, and consistent with the stance, but the justification is insufficiently explicit; and (5) the rationale explicitly and clearly links the FoM, the value definition, and the assigned stance, providing a complete and coherent explanation. We then computed the pairwise mean Quadratic-weighted Cohen's kappa (QWK) (Cohen, 1968) as 0.81, which is considered a strong level of agreement.
Following the pilot study, we consolidated and refined the annotation guidelines for annotating FoMs with the HVH using our HVD-HF framework. As illustrated in Figure 5, these guidelines establish clear procedures for verifying value paths, stances, and rationales, along with three explicit criteria for rationale revision. Annotators are also provided with comprehensive definitions of key concepts—misogyny, FoM, MP, and the full HVH—as well as illustrative examples demonstrating how an input FoM is annotated with corresponding (value, stance) paths and rationales.

4. Annotation results

After Phase A of the HVD-HF method, the LLM identified 3,377 Value Paths with a *support* stance and 7,480 value paths with an *ignore* stance. Of these annotated Value Paths (VPs), 591 VPs with*support* stance and 771 VPs *ignore* stance were flagged as low-confidence through self-consistency decoding. Annotators subsequently verified these low-confidence Value Paths. A low-confidence Value Path was accepted if a majority of annotators agreed the it was elicited by the FoM on which it is annotated. The stance of the accepted Value Path was determined by majority vote.
From the low-confidence set, annotators accepted 230 Value Paths with *support* stance and 269 Value Paths with *ignore* stance. They also revised the

Human Value Category	Support	Ignore
Tradition	604	12
Power (dominance)	568	4
Conformity (rules)	411	22
Face	275	6
Security (societal)	194	17
Hedonism	168	31
Conformity (interpersonal)	136	51
Security (personal)	80	58
Achievement	79	52
Power (resources)	70	1
Self-direction (action)	58	918
Universalism (concern)	48	1162
Benevolence (caring)	37	685
Universalism (objectivity)	32	430
Stimulation	32	32
Self-direction (thought)	16	898
Universalism (tolerance)	12	1007
Benevolence (dependability)	11	68
Humility	8	5
Universalism (nature)	0	0

Table 1: Distribution of Human Value Categories (Level 2 in the Human Value Hierarchy) across support or ignore stance values.

stance of 23 accepted Value Paths, changin the stance value from *ignore* to *support*. Similarly, and 11 Value Paths had their stance valeu changes from *support* to *ignore*. The resulting stance annotations of Value Paths (VPs) have a distributions of **3,051** VPs with support stance and **7,007** VPs with ignore stance. Additionally, out of the accepted Value Paths, the rationales of 193 VPs with a *support* stance and 167 VPs with an *ignore* stance were edited.

Inter-annotator agreement, measured by Krippendorff's α (Krippendorff, 2011), was 0.81 for annotating FoMs with value paths and 0.88 for assigning stances. For rationales, the one with the highest mean Likert rating among annotators was selected as the final rationale, with a pairwise mean Quadratic Weighted Cohen's κ (QWK) (Cohen, 1968) of 0.83 for these ratings.

5. Analysis of Annotated Human Values

A total of 1089 FoMs were annotated with Value Paths and their stance towards the FoM. For the support stance, higher-order values such as *Conservation* and *Self-Enhancement* dominate, whereas for the ignore stance, *Self-Transcendence* and *Openness to Change* are most prevalent. Table 1 illustrates the distribution of value categories across stances: categories such as *Tradition, Power (dominance)*, and *Conformity (rules)* are frequent in the support stance, while *Universalism (concern), Universalism (tolerance)*, and *Self-direction (action)* are more common in the ignore stance. Detailed statistical significance analyses supporting these observations are provided in Appendix D and additional dataset statistics are shown in Appendix C.

Support stance	Ignore Stance
Be respecting traditions	Have equality
Have the right to command	Be just
Be behaving properly	Be broadminded
Be compliant	Have freedom of thought
Have a stable society	Be independent
Have pleasure	Have freedom of action
Have a good reputation	Be loving
Have influence	Have the wisdom to accept others
Have social recognition	Be choosing own goals
Be polite	Have an objective view

Table 2: Top 10 human values (Level 1 in the Human Value Hierarchy) having support or ignore stance values.

All value categories are elicited by at least one FoM in either stance, with the exception of *Universalism (nature)*. At level 1 of the HVH, all but four values—*Be protecting the environment, Have harmony with nature, Have a world of beauty*, and *Be honouring elders.*—are elicited. The top 10 most frequent values for both stance values are listed in Table 2. The human values having a support stance indicate the most prevalent motivations of the meme creators, whereas the human values having an ignore stance value showcase which women's values are ignored by misogynist meme creators.

6. Using the MISOMEM-Val Dataset

We believe that the MISOMEM-Val dataset can be used in two ways. First it can be used for learning to automatically identify human values elicited by memes. Second, it can be used to generate counter-narratives that respond to misogynistic memes.

6.1. Learning to Automatically discern Human Values elicited by Memes

We formulate value identification as a multi-label classification task in which the model must predict all relevant human values (54 in total) associated with a given Frame of Misogyny (FoM), which is evoked by a meme. The training set consists of 829 (75%) FoMs, the validation set consists of 111 (10%) FoMs, and the test set consists of 165 (15%). We consider several methods for learning to recognize the human values. We consider a (**Majority**) method that assigns the five most frequent values to every test FoM, irrespective of content. For stance prediction, it always predicts the most common stance (support or ignore) associated with

System	Macro-F1 (Values)	Weighted-F1 (Values)	Macro-F1 (Stance)	Weighted-F1 (Stance)
Majority	0.18	0.28	0.50	0.54
LR	0.35	0.41	0.56	0.61
SVM	0.36	0.43	0.57	0.62
DeBERTa-B	0.55	0.63	0.65	0.70
RoBERTa-B	0.56	0.63	0.65	0.71
DeBERTa-H	0.60	0.68	**0.70**	**0.74**
RoBERTa-H	**0.62**	**0.69**	0.69	0.73
LLaMA (Zero-Shot)	0.74	0.79	0.82	0.85
LLaMA (Few-Shot)	**0.79**	**0.82**	**0.85**	**0.89**

Table 3: Macro and weighted F1-scores for the prediction of human values and stance on the test set.

System	Better	Equivalent	Worse
LLaMA Zero-Shot	8.12%	29.65%	62.23%
LLaMA Few-Shot	**10.75%**	**33.54%**	**55.71%**

Table 4: Comparison of the quality of rationales generated by the LLaMA LLM against gold rationales from our corpus.

each value. We also consider classifiers such as Logistic Regression (**LR**) (Cramer, 2004) and Support Vector Machine (**SVM**) (Cortes and Vapnik, 1995)—trained in a One-vs-Rest fashion with balanced class weights and per-label threshold tuning to predict the 54 values. For these methods, stance prediction is performed by a sequential setup wherein separate binary LR and SVM classifiers (support vs. ignore) are trained using TF–IDF features, augmented with features representing human values, thereby conditioning stance prediction on value relevance. To leverage the hierarchical structure of our value taxonomy, we further develop a shared-encoder multi-head architecture (**DeBERTa-H**). A single DeBERTa encoder—initialized from the ValueNet (Qiu) and ValueArg (Kiesel et al., 2022) checkpoints (Borenstein et al., 2025)—encodes each frame, while three output heads jointly predict higher-order value dimensions, value categories, and fine-grained values. Each head operates as a multi-label classifier trained with binary cross-entropy (BCE) loss. Stance prediction is subsequently performed using the same encoder, augmented with an additional binary classification head (support vs. ignore) trained with cross-entropy loss on (frame, value) pairs. An equivalent hierarchical configuration is applied to RoBERTa (**RoBERTa-H**).

We report Macro-F1 and Weighted-F1 scores for both human value and stance prediction, averaged across three random seeds. All models are trained or fine-tuned with default hyperparameters. Notably, stance F1 scores are computed conditioned on gold value labels, isolating stance prediction performance from errors in value identification.

In addition, we evaluate large language models—LLaMA 3.3–70B (Touvron et al., 2023)—under both zero-shot (**LLaMA Zero-Shot**) and few-shot (**LLaMA Few-Shot**) settings. In the zero-shot setup, the model is prompted with task descriptions, including definitions of FoMs and the Human Values Hierarchy (HVH) (Kiesel et al., 2022), and the required output format, asking it to identify all relevant *(value, stance)* pairs—out of the 54 fine-grained values—expressed in a given frame, along with short supporting rationales. The few-shot setup extends this prompt with five in-context examples illustrating correctly annotated *(value, stance)* pairs and their rationales. Chain-of-Thought prompting (Wei et al., 2022a) is employed in both settings to encourage explicit rationale generation.

To assess rationale quality, we combine human and automatic evaluations. For human evaluation, annotators compare each model-generated rationale with the corresponding gold rationale for a given *(value, stance)* pair, judging whether it is *better, worse*, or *equivalent* in quality (Weinzierl and Harabagiu, 2024a). For automatic evaluation, we compute BERTScore (Zhang et al., 2020) to measure semantic similarity and ROUGE-L (Lin, 2004) to capture lexical and structural overlap between generated and reference rationales.

Evaluation Results: Table 3 presents Macro-F1 and Weighted-F1 scores for human values and stance value prediction across the multiple methods we considered. Simple linear models (Logistic Regression and SVM) achieve modest performance, indicating that bag-of-words features capture some signal but are limited for the multi-label hierarchical value identification task. Transformer-based baselines DeBERTa-B and RoBERTa-B substantially improve performance, with hierarchical multi-head architectures (DeBERTa-H, RoBERTa-H) further boosting both values and stance scores, reflecting the benefit of modeling the three-level value structure jointly that allows parameter sharing and cross-level knowledge transfer. Notably, stance prediction consistently achieves higher F1-scores than predictions of human values, likely due to the simpler binary classification. Finally, large language models like LLaMA 3 show strong zero- and few-shot performance, outperforming traditional fine-tuned models, which highlights their ability to generalize and capture nuanced human values and stances even without task-specific training.

For rationale evaluation, the LLaMA few-shot setting attains slightly better semantic and lexical alignment—**ROUGE-L = 0.41, BERTScore = 0.74**—than the zero-shot variant **(ROUGE-L = 0.38, BERTScore = 0.70)**, confirming the benefit of in-context demonstrations. Human judgments align with these findings: following prior work that compares generated explanations against reference rationales using relative quality judgments (Weinzierl and Harabagiu, 2024a), few-shot rationales are more frequently rated as better or equivalent, whereas zero-shot rationales are more often considered worse relative to the gold references, as shown in Table 4.

6.2. Generating Counter Narratives

Counter narratives are commonly used as a strategy for combating online hate speech such as misogyny (Parker and Ruths, 2023). Excellent definitions of counter speech (or counter narratives) are provided in (Bonaldi et al., 2024), highlighting that counter narratives can be considered *"as non-aggressive textual feedback that uses credible evidence, factual arguments and alternative viewpoints*, cf. (Benesch, 2014; Schieb and Preuss, 2016). There are multiple tactics for generating counter-narratives, elaborated in (Bonaldi et al., 2024), including the generation of anti-stereotyping responses (Mun et al., 2023), or counter-argumentation generation as in (Albanyan et al., 2023).

The MISOMEM-Val dataset is enabling a different way of generating counter narratives, which is informed by Narrative Engagement Theory (Miller-Day and Hecht, 2013), which posits that compelling stories transport audiences, fostering deep emotional and cognitive immersion that drives persuasion and behavioral change. We have experimented with generating narratives, in the form of monologues or dialogues, in which Frames of Misogyny (FoMs) are referenced, as well as human values that they elicit or ignore. For this purposed, we have prompted ChatGPT. For example, when considering the FoM exemplified in Appendix A, namely *FoM: Women are single because they do not want a good man*, and selecting some of the value paths that are ignored by this FoM, (highlighted in red) we created the prompt for ChatGPT, illustrated in Figure 7.

Generate a monolog as a response to a misogynistic message that states "**women are single because they do not want good man**" you take into account human values such as **having freedom of thought**, and **women's goals of independence**

Figure 7: Prompt used for generating a monologue as a counter narrative relying on human values ignored by a Frame of Misogyny.

The counter narrative created by ChatGPT is illustrated in Appendix E. The narrative is persuasive, an important attribute of successful counter narratives, cf (Alyahya and Aldayel, 2024). Interestingly, the monologue also refers to additional values, e.g. mutual respect, shared responsibility, and genuine partnership, which do not belong to the HVH.

7. Conclusion

This paper introduced MISOMEM-Val, the first dataset that annotates human values in misogynistic memes through their underlying Frames of Misogyny (FoMs). By relying on the Human Value Hierarchy and the Taxonomy of Misogyny (ToM), MISOMEM-Val bridges the gap between human values theory and computational hate analysis while providing stance information capturing the relations between Frames of Misogyny and human values. Our analysis reveals that misogynistic frames predominantly support conservative human values, as well as power-related values and saving face. At the same time, they largely disregard women's values of self-direction and objectivity.

The annotation methodology used to create MISOMEM-Val further demonstrates how LLMs can be guided to reason hierarchically, generate interpretable rationales, and effectively complement human annotation. Finally, our experiments on automatically recognizing human values and their associated stances using MISOMEM-Val yielded promising results. We hope this work will encourage further research on developing computational models for automated human value recognition in misogynistic memes.

As future work, we plan to extend the HVA-HF annotation framework beyond revealing human values elicited by misogyny, annotating human values elicited by other forms of multimodal online hate, such as racism, homophobia, and ideological extremism. We also envision applying the annotation framework to broader multimodal content, including short-form videos and mixed-media social posts.

Limitations

Requirement for Available Frames

Our approach fundamentally depends on the availability of Frames of Misogyny (FoMs). The HVD-HF framework annotates human values only based on these FoMs, assuming that each meme evokes an articulated FoM. In the absence of such frames, the method cannot directly infer values from raw multimodal content. Consequently, the applicability of MISOMEM-Val is currently limited to datasets where frames have already been articulated. A natural direction for future work is the automatic discovery of frames directly from multimodal content, reducing the dependency on pre-existing taxonomies. Integrating frame induction methods with HVD-HF would enable end-to-end value annotation pipelines that generalize to new domains without manual frame construction. Such extensions would substantially improve the scalability of MISOMEM-Val-style annotations across emerging forms of online hate.

Annotation Bias

As with any human annotation effort, subjective interpretations may vary among annotators due to differences in cultural background, linguistic understanding, and personal beliefs (Chaturvedi et al., 2018). While the annotators differed in gender and nationality, the overall demographic diversity remains limited due to the small size of the annotation team, which may introduce shared biases in the perception of misogyny. To mitigate this, all LLM-generated value paths, stance assignments, and rationales were independently verified and edited by annotators following detailed guidelines, ensuring consistency and transparency.

Corpus Size

While 1089 Frames of Misogyny (FoMs) were richly annotated with 10,058 human value instances, the dataset remains relatively small compared to large-scale text corpora. This limited size reflects the complexity of the task, which demands high-quality, multi-level annotations grounded in human judgment. Nonetheless, our semi-automated LLM-assisted methodology enables scalable extension to broader domains in future work.

Use of LLMs

While our HVD-HF framework leverages large language models (LLMs) for value elicitation, their probabilistic nature may introduce biases or inconsistencies in initial outputs. Although self-consistency decoding and human verification substantially reduce such artifacts, minor variations in model responses can persist.

Ethical Statement

Our work adheres fully to the ACL Code of Ethics[2]. We utilize the publicly available MAMI (Multimedia Automatic Misogyny Identification) dataset released as part of SemEval-2022 Task 5, which was collected and distributed for research purposes. We follow the data usage conditions, licensing terms, and platform policies under which the benchmark was released, and do not perform additional data collection or redistribution beyond the scope permitted by the benchmark. Given the sensitive nature of this dataset, we implemented strict ethical safeguards to ensure that all stages of data handling and analysis were conducted responsibly. The study received approval from the Institutional Review Board (IRB) at the University of Texas at Dallas for the use of social media data. To reduce privacy risks, identifiable faces shown in illustrative figures are blurred, following common practice in prior work on multimodal hate speech.

To maintain the highest standards of reliability, we adopted a rigorous annotation protocol supported by inter-annotator agreement measures to verify consistency and quality. Because the dataset contains potentially harmful content, annotators were informed in advance about the nature of the material and participated voluntarily, with the option to discontinue at any time. In addition, our FoM-based annotation framework reduces direct exposure to raw multimodal content, since annotators primarily verify values at the framing level rather than repeatedly interpreting meme images. While we recognize the potential risks associated with engaging with offensive or harmful content, our goal is to advance the understanding of online misogyny and to promote the development of models and resources that mitigate its impact. Ultimately, this research contributes to both the scientific advancement of natural language processing and the broader pursuit of social justice and digital well-being.

Bibliographical References

Valuenet: A new dataset for human value driven dialogue system. 36.

Alan Agresti and Brent Coull. 1998. Approximate is Better than "Exact" for Interval Estimation of

[2] https://www.aclweb.org/portal/content/acl-code-ethics

Binomial Proportions. *The American Statistician*, 52(2):119–126.

Rakshitha Rao Ailneni and Sanda M. Harabagiu. 2025. Automatically discovering how misogyny is framed on social media. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12189–12208, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023. Not all counterhate tweets elicit the same replies: A fine-grained analysis. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 71–88, Toronto, Canada. Association for Computational Linguistics.

Ghadi Alyahya and Abeer Aldayel. 2024. Hatred stems from ignorance! distillation of the persuasion modes in countering conversational hate speech. In *International Conference on Web and Social Media*.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017, Seattle, United States. Association for Computational Linguistics.

Sandra J. Ball-Rokeach, Milton Rokeach, and Joel W. Grube. 1986. The great american values test : influencing behavior and belief through television.

Trevor J. M. Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse Processes*, 11(1):1–34.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

Nadav Borenstein, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025. Investigating human values in online communities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1607–1627, Albuquerque, New Mexico. Association for Computational Linguistics.

Duane Brown and R. Kelly Crace. 2002. *Life Values Inventory Facilitator's Guide*. Life Values Resources, Chapel Hill, NC.

Devichand Budagam, Ashutosh Kumar, Mahsa Khoshnoodi, Sankalp KJ, Vinija Jain, and Aman Chadha. 2025. Hierarchical prompting taxonomy: A universal evaluation framework for large language models aligned with human cognitive principles.

Iti Chaturvedi, Erik Cambria, Roy E. Welsch, and Francisco Herrera. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297.

J. S. Cramer. 2004. The early origins of the logit model. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 35(4):613–626.

John Duignan. 2016. Cramer's v.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communications*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 653–670, Online. Association for Computational Linguistics.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, José Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World values survey: Round seven - country-pooled datafile. Version 2.0.0, www.worldvaluessurvey.org.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saurav Joshi, Filip Ilievski, and Luca Luceri. 2023. Contextualizing internet memes across social media platforms. *Companion Proceedings of the ACM Web Conference 2024*.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhengyuan Liu, Yong Keong Yap, Hai Leong Chieu, and Nancy Chen. 2023. Guiding computational stance detection with expanded stance triangle framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3987–4001, Toronto, Canada. Association for Computational Linguistics.

Robert Lo, Abishek Sridhar, Frank Xu, Hao Zhu, and Shuyan Zhou. 2023. Hierarchical prompting assists large language model on web navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10217–10244, Singapore. Association for Computational Linguistics.

Michelle Miller-Day and Michael L. Hecht. 2013. Narrative means to preventative ends: A narrative engagement framework for designing prevention interventions. *Health Communication*, 28:657 – 670.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, MohammadAli Sadraei-Javaheri, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2024. The touché23-ValueEval dataset for identifying human values behind arguments. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134, Torino, Italia. ELRA and ICCL.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore. Association for Computational Linguistics.

Sara Parker and Derek A. Ruths. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Milton Rokeach. 1973. *The Nature of Human Values*. Free Press, New York.

Milton Rokeach. 1974. The nature of human values.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. pages 1–23, Fukuoka, Japan.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25:1–65.

Shalom H. Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50(4):19–45.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alberto Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kamile Demirutku, Oğuz Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.

Juan C. L. Teze, Antoni Perelló-Moragues, Lluis Godo, and Pablo Noriega. 2019. Practical reasoning using values: an argumentative approach based on a hierarchy of values. *Annals of Mathematics and Artificial Intelligence*, 87(3):293–319.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

T. L. van der Weide, F. Dignum, J. J. Ch. Meyer, H. Prakken, and G. A. W. Vreeswijk. 2010. Practical reasoning using values. In *Argumentation in Multi-Agent Systems*, pages 79–93, Berlin, Heidelberg. Springer Berlin Heidelberg.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Maxwell Weinzierl and Sanda Harabagiu. 2024a. Discovering and articulating frames of communication from social media using chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1617–1631, St. Julian's, Malta. Association for Computational Linguistics.

Maxwell A. Weinzierl and Sanda M. Harabagiu. 2024b. The impact of stance object type on the quality of stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15942–15954, Torino, Italia. ELRA and ICCL.

Kosuke Yoshimura and Hisashi Kashima. 2025. Hierarchical text classification using black box large language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A. Example of Memes, FoM and annotated value paths

Example of a Frame of Misogyny (FoM), the memes that evoke the FoM, and the annotated value paths with their corresponding stance toward the FoM are shown in Figure 8.

B. Prompting Details

We design the system prompt to guide the LLM toward hierarchical human value discovery consistent with the Human Value Hierarchy (HVH) and the Taxonomy of Misogyny (ToM). The prompt frames the model as an expert annotation assistant and provides explicit definitions of Frames of Misogyny (FoMs), Misogyny Problems (MPs), and the hierarchical organization of human values.

The system prompt instructs the model to perform value identification using a coarse-to-fine reasoning strategy. Specifically, the model is guided to first identify higher-order values (Level 3), followed by value categories (Level 2), and finally fine-grained human values (Level 1). This hierarchical decomposition encourages the model to generate coherent value paths aligned with the structure of the HVH rather than producing isolated value predictions.

In addition to value identification, the prompt requires the model to assign a stance to each value path. Two stance labels are defined: *support*, indicating that the FoM promotes or aligns with a value, and *ignore*, indicating that the FoM disregards or

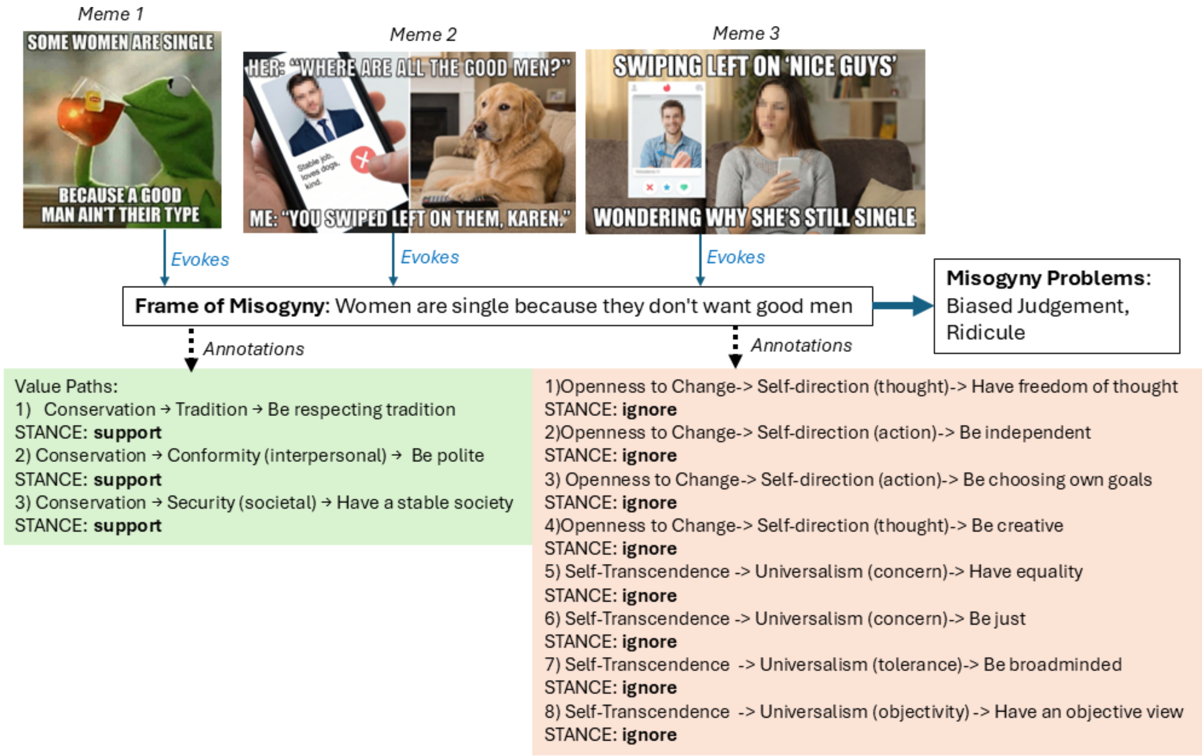

Figure 8: Example of a Frame of Misogyny (FoM), the memes that evoke the FoM, and the annotated value paths with their corresponding stance toward the FoM.

conflicts with the value. Stance is explicitly defined from the perspective of the frame itself. For each predicted value path, the model is further instructed to generate a rationale explaining both the value selection and the assigned stance, grounded in the definitions provided by the HVH and the associated MP.

To ensure structured outputs suitable for downstream processing, the prompt enforces hierarchical consistency across levels and constrains responses to a predefined JSON schema. The schema requires values, stances, and rationales to be produced in a machine-readable format, enabling automatic aggregation and subsequent human verification within the HVD-HF framework.

C. Additional Dataset Statistics

To examine how human values are represented across the full MISOMEM corpus, we computed the frequency of unique frames eliciting each human value (from Level 1 of the Human Values Hierarchy (HVH)). Figure 10 shows the distribution of human values ranked by the number of Frames of Misogyny (FoMs) that elicit them. The distribution is strongly skewed. A small set of values accounts for a large proportion of the dataset, with *Be respecting traditions* and *Have the right to command* appearing substantially more frequently than other values. These are followed by values related to conformity and social order, including *Be behaving properly*, *Be compliant*, and *Have a stable society*. Mid-frequency values include reputation, influence, and social recognition, while a long tail of values appears only rarely. These low-frequency values are primarily individual-oriented or abstract in nature, such as curiosity, creativity, and self-direction, indicating that personal autonomy and exploratory values are comparatively underrepresented within misogyny frames.

To further understand how values vary across different misogyny problem types, we computed the frequencies of value categories for each problem category (Figure 11). This analysis reveals several consistent patterns. Traditional and power-related categories, including Tradition and Power (dominance), appear across nearly all misogyny problem types and are particularly prominent in categories such as Patriarchal attitudes, Objectification, and Stereotyping. Conformity-related values also show high counts, explaining that these values are broadly associated with misogynistic discourse across contexts.

D. Statistical Analysis of Human Value Prevalence

To provide statistical support for the prevalence patterns discussed in Section 5, we conduct additional analyses on both value categories (Level-2)

Human Value Category	Support	Ignore	Total	Prev.	95% CI
Universalism (concern)	48	1162	1210	14.6%	[13.8,15.3]
Universalism (tolerance)	12	1007	1019	12.3%	[11.6,13.0]
Self-direction (action)	58	918	976	11.8%	[11.1,12.5]
Self-direction (thought)	16	898	914	11.0%	[10.3,11.7]
Benevolence (caring)	37	685	722	8.7%	[8.1,9.3]
Tradition	604	12	616	7.4%	[6.9,8.0]
Power (dominance)	568	4	572	6.9%	[6.3,7.4]
Universalism (objectivity)	32	430	462	5.6%	[5.1,6.1]
Conformity (rules)	411	22	433	5.2%	[4.7,5.7]
Face	275	6	281	3.4%	[3.0,3.8]

Table 5: Prevalence of the most frequent Human Value Categories.

and higher-order values (Level-3) from the Human Value Hierarchy. The counts used in this analysis correspond to the support and ignore distributions reported in Table 1.

Across the dataset, the total number of value category annotations is $N = 8,298$, obtained by summing the support $(2,839)$ and ignore $(5,459)$ stance counts.

D.1. Prevalence Estimation

Higher-Order Value	Prevalence	95% CI
Self-Transcendence	42.1%	[41.0, 43.2]
Openness to Change	15.1%	[14.3, 15.9]
Conservation	13.9%	[13.1, 14.7]
Self-Enhancement	8.5%	[7.9, 9.1]

Table 6: Prevalence of higher-order human values.

Table 5 reports prevalence estimates for the most frequent value categories.

The prevalence of a value group v_i is defined as:

$$P(v_i) = \frac{n_i}{N} \tag{1}$$

where n_i denotes the number of annotations associated with value group v_i, and N is the total number of value annotations.

To quantify uncertainty, we compute 95% confidence intervals using the binomial proportion approximation (Agresti and Coull, 1998):

$$CI = p \pm 1.96\sqrt{\frac{p(1-p)}{N}} \tag{2}$$

where p denotes the observed prevalence.

Prevalence of higher-order human values is shown in Table 6.

D.2. Association Between Value Categories and Stance

To evaluate whether value categories are associated with stance assignments, we conduct a chi-square test of independence (Pearson, 1900). The analysis yields:

$$\chi^2(18) = 6480.79, \quad p < 0.001 \tag{3}$$

indicating a statistically significant association between value categories and stance.

To quantify the strength of this relationship, we compute Cramér's V (Duignan, 2016):

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{4}$$

where χ^2 denotes the chi-square statistic, N is the total number of observations in the contingency table ($N = 8{,}298$), and k is the smaller dimension of the table. Since the contingency table contains 19 value categories (Universalism–nature has zero observations and is excluded) and 2 stance labels, $k = 2$. The resulting value is:

$$V = 0.884 \tag{5}$$

which indicates a very strong association between value categories and stance assignments.

E. Counter Narrative for Misogyny

In response top the prompt from Figure 7, ChatGPT generated the following monologue, in which: (a) citations from the Frame of Misogyny are given; (b) reference to ignore human values are present (underlined). The narrative also provides arguments and explanations. The counter narrative is:

People often say that women are single because they "don't want good men," but that statement assumes that every woman's goal should be to be in a relationship and that being single must mean something is wrong with her choices. In reality, people's lives are much more complex than that.

Women, like anyone else, have the freedom to decide what they want for their lives. Some women focus on education, careers, friendships, family responsibilities, or personal growth. Others may want relationships but are *thoughtful about choosing partners who respect them, share their values, and support their goals. Wanting equality, kindness, and mutual respect in a relationship does not mean rejecting "good men"; it means valuing healthy partnerships.*

Being single is not a failure or a problem to explain away. It can simply be a life stage or even a deliberate choice. Independence allows people to understand themselves, pursue ambitions, and build meaningful lives. A relationship should add to someone's well-being, not replace their autonomy or limit their opportunities.

Human relationships work best when both people recognize each other's dignity, freedom of thought, and personal aspirations. Women today often have opportunities that previous generations fought hard to secure: education, careers, financial independence, and the ability to make their own decisions. Choosing how and when to share one's life with a partner is part of that freedom.

So instead of assuming that women are single because they reject "good men," it might be more accurate to recognize that many women simply expect relationships built on mutual respect, shared responsibility, and genuine partnership. Those values benefit everyone, not just women.

```
You are an expert annotation assistant for identifying human
values and their stance in Frames of Misogyny (FoMs).
A Frame of Misogyny highlights specific aspects of misogyny
by defining a problem, implying causes, evoking moral or
human values, and suggesting implicit or explicit evaluations
or solutions.
Your task is to identify HUMAN VALUE PATHS elicited by a
given Frame of Misogyny using the Human Value Hierarchy
(HVH), and determine the stance of the frame toward each
value.
-------------------------------------
HUMAN VALUE HIERARCHY (HVH)
-------------------------------------
The HVH is organized hierarchically:
- Level 3: Higher-order values (broad dimensions)
- Level 2: Value categories
- Level 1: Fine-grained human values
Each Level-1 value belongs to a Level-2 category, and each
Level-2 category belongs to a Level-3 value.
Value Taxonomy:
{values_yaml}
-------------------------------------
STANCE DEFINITION
-------------------------------------
For every identified value, assign ONE stance:
- support: the Frame of Misogyny promotes, endorses, or
aligns with the value.
- ignore: the Frame of Misogyny disregards, suppresses, or
conflicts with the value.
Stance is determined from the perspective of the frame itself
(i.e., what the frame promotes or ignores).
-------------------------------------
TASK INSTRUCTIONS
-------------------------------------
Use hierarchical reasoning to discover value paths and
stance.
Follow a coarse-to-fine process:
Step 1:  Identify Level-3 values
Determine which higher-order values are relevant to the FoM.
Step 2: Identify Level-2 values
For each Level-3 value, identify compatible Level-2
categories.
Step 3: Identify Level-1 values
For each Level-2 category, identify specific Level-1 values
elicited by the frame.
For EACH value path:
1. Output the value at each level.
2. Assign a stance (support or ignore).
3. Provide a rationale explaining:
  - why the value is elicited by the FoM
  - why the assigned stance is appropriate
  - how this relates to the Misogyny Problem definition.
-------------------------------------
IMPORTANT CONSTRAINTS
-------------------------------------
- Follow the hierarchy strictly: Level-3 → Level-2 → Level-1.
- Ensure consistency across levels (children must match
parent values).
- Do NOT output isolated values without hierarchical
alignment.
- Stance must be assigned to every value path.
- Select only values clearly grounded in the frame.
- Keep rationales analytical, concise, and objective.
- Do not include speculative interpretations.
Return results strictly following the required JSON schema.
```

Figure 9: System prompt design used in the HVD-HF framework.

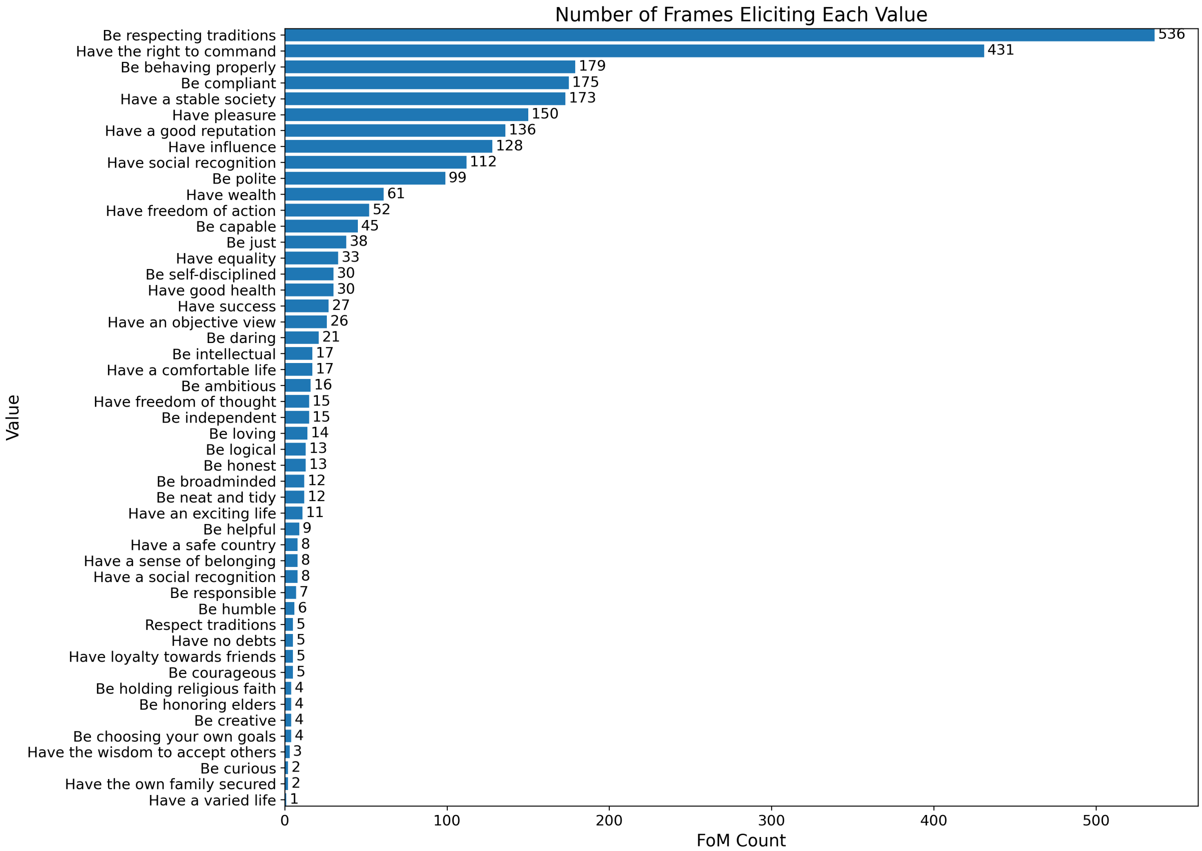

Figure 10: Number of FoMs Eliciting each Human Value (Level 1 of the Human Value Hierarchy).

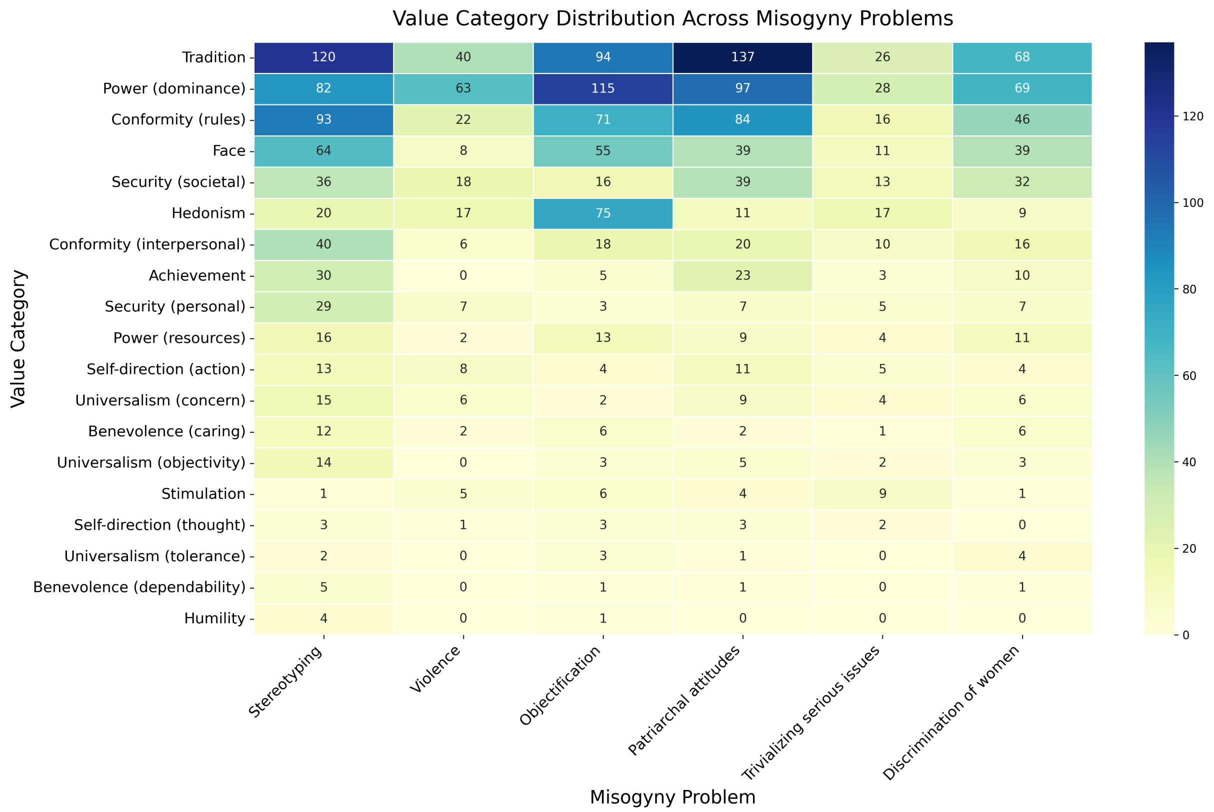

Figure 11: Distribution of Human Value Categories (Level 2 of the Human Value Hierarchy) across Misogyny Problems (available from the Taxonomy of Misogyny).