Social Perception as Theory of Mind: Evidence from Annotator Disagreement in Sexism Detection

Diana Nurbakova

INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205
69100 Villeurbanne, France
diana.nurbakova@insa-lyon.fr

Abstract

Social perception, i.e., how people form impressions from language, forms the basis for subjective NLP tasks like sexism detection. Yet the mechanisms driving annotator disagreement remain unexplained. We propose that Theory of Mind (ToM) provides this mechanism: annotators perform cognitive ToM (recognising norm-relevant content and inferring speaker intent) and affective ToM (estimating target impact), producing disagreement when these inferences diverge. Analysing the EXIST 2025 dataset (7,958 tweets, 4,044 memes), we find a detection-interpretation dissociation. Annotators who agree content is sexist show higher disagreement about speaker intent and sexism type than about detection itself. This gap replicates across tweets and memes modalities. Perceiver gender does not affect detection for text but does for memes (where even detection is ToM-demanding), shows negligible effects on intent attribution, and selectively shapes harm-related categorisation (*misogyny* in text, *objectification* in memes) while leaving abstract categories unaffected. Gender structuring thus increases with affective ToM demand. These patterns replicate established social psychology findings at computational scale and demonstrate that social perception operates through dissociable ToM processes that current NLP systems collapse.

Code — https://github.com/diana-nurbakova/ToM-on-sexism-data

Datasets — https://nlp.uned.es/exist2025/

Introduction

Social perception, the process by which people form judgments about speakers, targets, and social situations from linguistic and visual cues, is central to NLP tasks such as sexism detection, hate speech classification, and toxicity assessment. Most NLP systems treat these socially grounded judgments as fixed properties of text, collapsing the variability of human perception into single gold-standard labels (Basile et al. 2021).

The *Learning with Disagreement (LeWiDi)* paradigm has shown that annotator disagreement in subjective social tasks is pervasive and structured, not noise (Plank 2022; Leonardelli et al. 2023; Rottger et al. 2022; Alacam et al. 2025). The field lacks, however, an *explanatory* account. Why do annotators disagree? Why do they disagree in systematic patterns?

We propose to use **Theory of Mind** (ToM), the capacity to attribute mental states to others (Premack and Woodruff 1978), as the cognitive mechanism underlying social perception in language. Judging whether a text is sexist requires attributing communicative intent to the speaker, estimating impact on the target, and evaluating social norm violations. These are ToM operations. Different perceivers bring different priors to them, producing structured disagreement.

Using the EXIST 2025 sexism detection dataset (Plaza et al. 2026), we provide three contributions: (1) a theoretical framework mapping annotation tasks to specific ToM component processes; (2) empirical evidence of a detection-interpretation dissociation consistent with the cognitive-affective ToM distinction; (3) cross-modal replication across tweets and memes showing that perceiver gender selectively shapes affective-ToM-dependent harm categories. All analysis scripts and figures are available at our GitHub repository.

Theory of Mind as the Mechanism of Social Perception

ToM Component Processes

Theory of Mind is the capacity to attribute mental states (beliefs, desires, intentions, emotions) to others and to understand that these states may differ from one's own (Premack and Woodruff 1978). ToM is not a monolithic ability. It comprises dissociable component processes: agent identification, belief tracking, intent attribution, and affective state inference (Schaafsma et al. 2015).

We distinguish **cognitive ToM**, i.e., inferring others' beliefs and intentions, from **affective ToM**, i.e., inferring others' emotional states. The study by Shamay-Tsoory and Aharon-Peretz (2007) suggests that cognitive ToM is a prerequisite for affective ToM. Affective ToM additionally is closely related to empathic processes (Shamay-Tsoory 2011), which depends on the perceiver's capacity to simulate the target's emotional experience. Since perspective-taking is anchored in one's own experience (Epley et al. 2004), perceivers with greater experiential proximity to the target's situation will produce different affective inferences. We use the

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

term *ToM demand* (Apperly 2012) to characterise the complexity of the mental state inferences a task requires.

Annotation as Implicit ToM

The EXIST 2025 annotation guidelines (Plaza et al. 2026) operationalise a gradient of ToM demand across three tasks (Figure 1) without explicitly stating it.

Task 1: Sexism identification. Annotators judge whether a text "expresses sexist ideas because it is sexist itself, describes a sexist situation, or criticizes a sexist behavior." It is a binary classification problem with YES/NO labels. This does not require judging that the speaker is sexist. It requires recognising that the text engages with a gender norm, whether by violating it (DIRECT), reporting a violation (REPORTED), or condemning one (JUDGEMENTAL). Mind that identifying this second-level sexist category is the goal of Task 2. We associate this with *norm-relevance recognition*: a cognitive ToM operation at its lowest demand level, identifying that the content concerns a social norm about gender, regardless of the speaker's stance toward it.

Task 2: Source intention. Annotators classify the speaker's communicative goal as DIRECT (intending sexism), REPORTED (reporting sexism), or JUDGEMENTAL (condemning sexism). These categories are mutually exclusive. This is explicit intent attribution, a core cognitive ToM operation. Distinguishing reporting from condemning requires modelling layered communicative intentions, engaging higher-order ToM (Perner and Wimmer 1985). Consider the contrast. A DIRECT tweet like "*A woman needs love, to fill the fridge [...] I don't see what else she needs*" wears its intent on the surface. A JUDGEMENTAL tweet like "*As usual, the woman was the one quitting her job for the family's welfare...*" demands detecting sarcasm in "*as usual*", attributing a critical stance to the speaker, and recognising an implicit claim about gendered expectations. For memes, only two categories are used: DIRECT/JUDGEMENTAL.

Task 3: Sexism categorization. Annotators assign one or more of five categories depending on the focus of sexist attitudes: ideological-inequality, stereotyping-dominance, objectification, sexual-violence, misogyny. These categories differ in ToM demand. *Ideological-inequality* and *stereotyping-dominance* involve recognising abstract propositional content about gender (e.g., "women are inferior," "women belong in the home"): a cognitive ToM operation. *Objectification*, *sexual-violence*, and *misogyny* involve estimating harm directed at persons and bodies, which requires simulating the target's experience: an affective ToM operation (Shamay-Tsoory and Aharon-Peretz 2007).

The guidelines never ask annotators to match surface-level lexical patterns. They ask them to infer intent, identify the facets being the focus of sexist attitude, and integrate both into a judgment. The annotation scheme assumes ToM. We make this assumption explicit.

Predictions

Our study tests the following three predictions.

P1 (Dissociation): Disagreement should be substantially higher at the interpretation level than at the detection level, since interpretation requires more complex ToM inferences. The cognitive-affective distinction predicts a *gap* between levels, not that detection will show no disagreement. It. A noise account predicts correlated disagreement. The ToM account predicts a dissociation.

P2 (Perceiver modulation): Gender should affect perception at the level where affective ToM is engaged, i.e., where experiential proximity matters. For text, where detection cues are largely verbal and explicit, this means gender should shape interpretation more than detection. For multimodal content, where even detection requires interpreting visual irony and implicit framing, gender effects may emerge earlier (Swim and Cohen 1997).

P3 (Selective categories): The gender effect should concentrate on categories that require estimating bodily or emotional harm to the target (sexual violence, misogyny, objectification), i.e., categories that engage empathic simulation. Categories involving abstract propositional content (ideological inequality, stereotyping) should show no gender effect, consistent with the cognitive-affective dissociation (Shamay-Tsoory and Aharon-Peretz 2007).

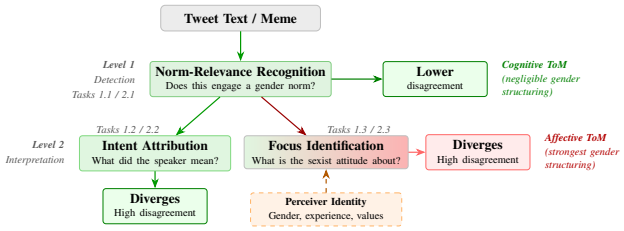

Figure 1: Two-level model of social perception as ToM. Green = cognitive ToM (norm-relevance recognition, intent attribution); red = affective ToM. Focus identification (Task 3) receives a gradient fill because it engages both cognitive ToM (recognising abstract propositions) and affective ToM (identifying harm to persons). Both Level 2 operations produce high disagreement, but only the affective component is modulated by perceiver identity (dashed arrow).

Experiments

Data

We analyse the EXIST 2025 dataset (Plaza et al. 2026), a third-party annotated resource (we did not perform any annotation ourselves): 7,958 tweets (EN: 3,749; ES: 4,209) and 4,044 memes (EN: 2,010; ES: 2,034). The dataset contains two modalities (tweets, memes), each annotated along three subtasks: sexism identification, source intention, and categorisation. EXIST numbers tweet subtasks as 1.1/1.2/1.3 and meme subtasks as 2.1/2.2/2.3; the definitions are otherwise identical across modalities. Each instance is annotated by 6 annotators (3 female, 3 male) sampled from a pool of 887 unique annotators recruited via Prolific; the specific 6 vary across instances. Annotator demographic metadata (gender, age, education, country) is available. Source intention is ternary for tweets (DIRECT/REPORTED/JUDGEMENTAL) and binary for memes (DIRECT/JUDGEMEN-

TAL). Categorisation is a 5-class multi-label task for both modalities. We measure disagreement using Shannon entropy for binary and categorical distributions, and mean pairwise Jaccard similarity for multi-label categorisation.

Detection–Interpretation Dissociation (P1)

Disagreement on detection is pervasive. It occurs in 67.1% of tweets and 76.3% of memes. Detection is not easy. This is expected: the EXIST detection task is deliberately broad, encompassing content that perpetrates, reports, or condemns sexism. High disagreement on such a task is the phenomenon the LeWiDi paradigm (Basile et al. 2021; Rottger et al. 2022) treats as signal rather than noise. But the critical finding is that interpretation-level disagreement is far higher still. Among 3,152 tweets where the majority labels the content as sexist, intent attribution entropy reaches 0.876, where 1.0 is the ternary maximum. Categorisation shows low overlap: mean Jaccard = 0.382. The pattern replicates for memes (n = 2,038 majority-YES): intent entropy = 0.701 (high for a binary distribution), Jaccard = 0.418 (Table 1). The key finding is the *gap*: the same annotators who disagree on 67% of detection judgments show near-total disagreement on what the speaker intended and what kind of sexism it is.

Consider the following example. For the tweet *"this is the same little boy who has said 'when we get married, I'm going to slap the shit out of you — fucking stupid bitch' "*, all six annotators agree the content is sexist. Yet they assign the full spectrum of intents: two say JUDGEMENTAL (the speaker is condemning the boy's behaviour), two say REPORTED (the speaker is documenting it), and two say DIRECT (the speaker is endorsing it). The intent entropy is maximal: the three categories are equally represented. On categorisation, each annotator selects a different subset (Jaccard = 0.189): one assigns MISOGYNY alone, another STEREOTYPING + OBJECTIFICATION, another SEXUAL-VIOLENCE alone. Is this a report of domestic violence socialisation, a condemnation, or an endorsement? Each reading requires a different model of speaker intent. Same text, same detection, radically different inference.

A noise account predicts correlated disagreement across levels: if labels were random, both detection and interpretation should show similar entropy. A guideline-ambiguity account predicts disagreement where guidelines are vague, but the binary detection task has the simplest guidelines and still shows 67% disagreement. The ToM account predicts this specific pattern: low-demand norm-relevance recognition produces lower disagreement than high-demand intent attribution and focus identification. Note that the interpretation-level metrics are computed among majority-YES instances, i.e., cases where annotators already partially agreed on detection. Conditioning on partial agreement should, if anything, select for more "obvious" sexist instances where interpretation might also be easier. That this subsample nonetheless shows interpretation disagreement approaching the theoretical ceiling. This strengthens the dissociation finding.

Gender Selectively Shapes Perception (P2 & P3)

Detection level. For tweets, female and male annotators show no significant difference in detection rates (F: 0.450, M: 0.461; Wilcoxon p = 0.142, r = −0.026). Of 960 perfect 3-3 splits, only 10.1% align along gender lines. At the detection level, disagreement is not driven by gender group membership. Mind that other demographic and ideological factors, which we did not test, may still structure it.

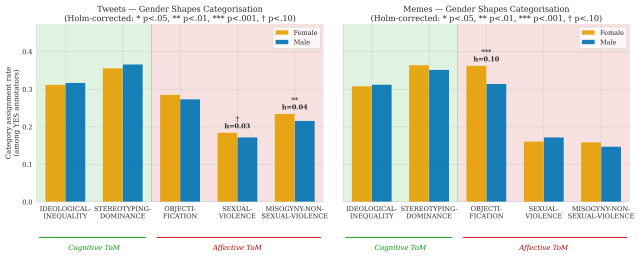

Figure 2: Gender shapes categorisation selectively. Left: tweets; right: memes. Significance markers use Holm-corrected p-values († $p_{\text{adj}} < .10$; $**$ $p_{\text{adj}} < .01$; $***$ $p_{\text{adj}} < .001$). Gender differences appear only for affective-ToM categories, never for cognitive-ToM categories.

For memes, women detect significantly more sexism (F: 0.586, M: 0.528; $p < 10^{-24}$, r = 0.24). The effect is small (5.8 percentage points) despite the extreme p-value, which reflects the large sample. This is consistent with P2's prediction that gender effects emerge where affective ToM is engaged. Memes encode sexism through visual framing, irony, and text-image incongruity. Even detecting sexism in this format requires interpretive work that pushes ToM demand higher than for text. Social psychology distinguishes *overt* from *subtle* sexism (Swim and Cohen 1997). drawing attention to subtle everyday sexism affects women's beliefs more than men's (Becker and Swim 2011), consistent with gender-differentiated processing of ambiguous content.

Intent level. For tweets, a small gender asymmetry emerges in intent attribution: women assign DIRECT (intending sexism) slightly more often, while men favour JUDGEMENTAL (condemning sexism) (h = 0.048, p_{adj} = 0.001; Table 1). REPORTED shows no gender difference. For memes, no gender effect on intent reaches significance. The intent effect sizes ($|h| < 0.05$) are smaller than even the weakest surviving categorisation effect (h = 0.045 for *misogyny*), suggesting that gender structuring increases with affective ToM demand rather than appearing as a sharp binary at the cognitive-affective boundary.

Categorisation level. Gender selectively shapes harm-related categories while leaving cognitive categories unaffected (Table 1; Figure 2). All per-category tests are corrected for multiple comparisons using the Holm-Bonferroni procedure within each modality. For tweets, only *misogyny* survives correction (p_{adj} = 0.005, h = 0.045, OR = 1.11). *Sexual violence* and *objectification* show trends in the same direction but do not reach significance after correction. For memes, only *objectification* survives ($p_{\text{adj}} < 10^{-8}$, h = 0.10, OR = 1.24). In both modalities, ideological-inequality and stereotyping-dominance show no gender effect.

This selectivity confirms the cognitive-affective split de-

Measure	Tweets	Memes
N instances	7,958	4,044
Any disagreement	67.1%	76.3%
Among majority-YES instances:		
N majority-YES	3,152	2,038
Detect. entropy	0.538	0.555
Intent entropy	0.876	0.701
Categ. Jaccard	0.382	0.418
Gender effect on detection:		
p-value (r)	0.142 (-0.03)	<0.001 (0.24)
Gender on intent (p / p_{adj}, h):		
Direct	**.008** / .015 (.04)	.14 / .29 (n.s.)
Reported	.57 / .57 (n.s.)	—
Judgemental	<**.001** / .001 (.05)	.14 / .29 (n.s.)
Gender on categ. (p / p_{adj}, h):		
Ideological-ineq.	.42 / .42 (n.s.)	.60 / .60 (n.s.)
Stereotyp.-domin.	.11 / .23 (n.s.)	.12 / .27 (n.s.)
Objectification	.053 / .16 (n.s.)	<**.001** / <**.001** (.10)
Sexual violence	.016 / .064 (n.s.)	.091 / .27 (n.s.)
Misogyny	**.001** / **.005** (.045)	.062 / .25 (n.s.)

Table 1: Cross-modal comparison. Intent and categorisation rows report raw p / Holm-corrected p_{adj}; bold = $p < .05$; h = Cohen's h. Gender effects on intent are negligible.

scribed in Section : the only two categories surviving correction across either modality are both affective-ToM categories, while neither cognitive-ToM category approaches significance. Although the individual effect sizes are small ($h < 0.11$), the pattern is consistent. Women are disproportionate targets of misogyny and visual objectification (Fredrickson and Roberts 1997; Swim et al. 2001). Their experiential proximity to these forms of harm produces different impact estimations through empathic simulation (Shamay-Tsoory and Aharon-Peretz 2007).

We note the modality-specific pattern. In text, where harm is described verbally, the surviving gender effect is *misogyny*. In memes, where harm operates through visual composition and the objectifying gaze (Mulvey 1975; Fredrickson and Roberts 1997), it is *objectification*. The mechanism is the same (affective ToM modulated by experiential proximity) but it expresses through each medium's affordances.

One meme example illustrates the gender split on categorisation (see Figure 3 in Appendix A). The meme reads: *"*Female character breathes* *Rule 34 artists* HUMAN ORGAN I CAN MAKE THIS WORK." Rule 34* is an internet norm that any fictional character will be sexualised. The meme satirizes this by reducing a female character to body parts. All three female annotators assigned OBJECTIFICATION, while no male annotator who detected sexism did. This asymmetry is consistent with women's experiential proximity to sexual objectification activating a different affective ToM inference about target impact.

Discussion

A Two-Level Model. Our findings support a two-level model of social perception. **Level 1** (detection) engages norm-relevance recognition; disagreement is already substantial (67% of tweets, 76% of memes), with gender effects absent for text but present (though small, $r = 0.24$) for multimodal content. **Level 2** (interpretation) engages intent attribution and focus identification. Both produce high disagreement, but gender structures them differently. Intent attribution shows negligible gender effects ($|h| < 0.05$). Harm categorisation shows larger effects (h up to 0.10), modulated by experiential proximity through empathic simulation. Gender structuring thus increases with affective ToM demand: absent or negligible for cognitive operations, present for affective ones. Current NLP systems often collapse perceiver variation into single labels.

Bridging Social Psychology and NLP. Our computational findings replicate patterns from controlled psychology experiments at scale. Glick and Fiske's (1996) ambivalent sexism framework distinguishes *hostile* (overt) from *benevolent* (subtle) sexism, predicting different psychological responses to each. We find a parallel in the detection-interpretation gap. Diary studies show that women report more sexist incidents than men, and these incidents affect their psychological wellbeing (Swim et al. 2001). We observe a related asymmetry at the detection-categorisation boundary. Value-based standards for judging sexism (Swim et al. 2005) predict that perceivers use their own experiential standards, consistent with our finding that gender selectively shapes harm categories. Objectification theory (Fredrickson and Roberts 1997) predicts heightened female sensitivity to visual objectification; the meme objectification effect ($p_{\text{adj}} < 10^{-8}$, $h = 0.10$) confirms this at scale, albeit with a small effect size characteristic of data.

Implications for NLP. The LeWiDi paradigm correctly preserves disagreement. Our analysis explains *why* it occurs. First, annotation protocols should capture not just labels but reasoning: what intent did the annotator attribute? What impact did they estimate? Separating these components would make the ToM structure of disagreement explicit and enable perceiver-aware modelling. Second, perceiver-aware approaches that condition on annotator identity (e.g., Pastells et al. 2025) should condition selectively: perceiver identity matters most for affective-ToM-dependent judgments and negligibly for cognitive ones. Third, soft metrics like the Information Contrast Model (Amigo and Delgado 2022) capture disagreement distributions but do not explain their structure. ToM-informed evaluation could distinguish perception-driven disagreement from noise.

Conclusion

Social perception is not a fixed property of text but a ToM-mediated inference at dissociable levels. When annotators agree a text is sexist but disagree about intent and focus, they reveal a gap between detection and interpretation that grows as ToM demand increases. Perceiver identity selectively shapes the affective layer: gender structuring is negligible for cognitive operations (detection, intent attribution, abstract categories) and strongest for affective operations (harm categorisation). Making the ToM structure of annotation explicit opens new directions for perceiver-aware NLP.

Limitations

We study a single task domain (sexism). Generalisation to other social perception tasks remains to be tested. We cannot fully rule out guideline ambiguity, though the selective gender effect on specific categories is not predicted by guideline accounts. Part of the interpretation-level disagreement may also reflect genuinely ambiguous class boundaries in the EXIST annotation scheme rather than divergent ToM inferences. Thus, disentangling task complexity from ToM demand is a direction for future work. Our mapping of categories to cognitive vs. affective ToM is necessarily coarse: cognitive ToM likely contributes to impact estimation as well (understanding the target's situation is a prerequisite for simulating their experience), but we argue the additional affective component is what produces gender-structured disagreement. Our analysis is observational: we infer ToM from annotation patterns rather than measuring cognitive processes directly. Our Fisher's exact tests treat each annotation as independent, but the same annotators appear across multiple instances, which inflates significance for category-level tests; multilevel models would be more appropriate for a full analysis. Gender is treated as binary, and annotator pools skew toward specific countries. The tweet and meme annotator pools overlap substantially (Jaccard = 0.91; matched on gender, age, and education), so cross-modal differences are unlikely to reflect population confounds, though modest differences in country composition exist (Cramér's V = 0.25). All observed gender effects on categorisation are small ($h < 0.11$). The pattern's theoretical significance lies in its selectivity rather than its magnitude.

References

Alacam, O.; Hoeken, S.; Säuberli, A.; Gröner, H.; Frassinelli, D.; Zarrieß, S.; and Plank, B. 2025. Disentangling Subjectivity and Uncertainty for Hate Speech Annotation and Modeling using Gaze. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 28695–28712. Suzhou, China: Association for Computational Linguistics.

Amigo, E.; and Delgado, A. 2022. Evaluating Extreme Hierarchical Multi-label Classification. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5809–5819. Dublin, Ireland: Association for Computational Linguistics.

Apperly, I. A. 2012. What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, 65(5): 825–839.

Basile, V.; Fell, M.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M.; and Uma, A. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, 15–21. Online: Association for Computational Linguistics.

Becker, J. C.; and Swim, J. K. 2011. Seeing the Unseen: Attention to Daily Encounters With Sexism as Way to Reduce Sexist Beliefs. *Psychology of Women Quarterly*, 35(2): 227–242.

Epley, N.; Keysar, B.; Van Boven, L.; and Gilovich, T. 2004. Perspective Taking as Egocentric Anchoring and Adjustment. *Journal of Personality and Social Psychology*, 87(3): 327–339.

Fredrickson, B. L.; and Roberts, T.-A. 1997. Objectification Theory: Toward Understanding Women's Lived Experiences and Mental Health Risks. *Psychology of Women Quarterly*, 21(2): 173–206.

Glick, P.; and Fiske, S. T. 1996. The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3): 491–512.

Leonardelli, E.; Abercrombie, G.; Almanea, D.; Basile, V.; Fornaciari, T.; Plank, B.; Rieser, V.; Uma, A.; and Poesio, M. 2023. SemEval-2023 Task 11: Learning with Disagreements (LeWiDi). In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2304–2318. Toronto, Canada: Association for Computational Linguistics.

Mulvey, L. 1975. Visual Pleasure and Narrative Cinema. *Screen*, 16(3): 6–18.

Pastells, P.; V{\'{a}}zquez Chas, M.; Farr{\'{u}}s, M.; and Taul{\'{e}}, M. 2025. CLiC at EXIST 2025: Combining Fine-tuning and Prompting with Learning with Disagreement for Sexism Detection. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025*, volume 4038 of *CEUR Workshop Proceedings*, 2100–2111. CEUR-WS.org.

Perner, J.; and Wimmer, H. 1985. "John thinks that Mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3): 437–471.

Plank, B. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10671–10682. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Plaza, L.; Carrillo-de Albornoz, J.; Arcos, I.; Rosso, P.; Spina, D.; Amigó, E.; Gonzalo, J.; and Morante, R. 2026. Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos. In Carrillo-de Albornoz, J.; García Seco De Herrera, A.; Gonzalo, J.; Plaza, L.; Mothe, J.; Piroi, F.; Rosso, P.; Spina, D.; Faggioli, G.; and Ferro, N., eds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 16089, 266–289. Cham: Springer Nature Switzerland. ISBN 9783032043535 9783032043542.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526.

Rottger, P.; Vidgen, B.; Hovy, D.; and Pierrehumbert, J. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–

190. Seattle, United States: Association for Computational Linguistics.

Schaafsma, S. M.; Pfaff, D. W.; Spunt, R. P.; and Adolphs, R. 2015. Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2): 65–72.

Shamay-Tsoory, S. G. 2011. The Neural Bases for Empathy. *The Neuroscientist*, 17(1): 18–24.

Shamay-Tsoory, S. G.; and Aharon-Peretz, J. 2007. Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia*, 45(13): 3054–3067.

Swim, J. K.; and Cohen, L. L. 1997. Overt, Covert, And Subtle Sexism: A Comparison Between the Attitudes Toward Women and Modern Sexism Scales. *Psychology of Women Quarterly*, 21(1): 103–118.

Swim, J. K.; Hyers, L. L.; Cohen, L. L.; and Ferguson, M. J. 2001. Everyday Sexism: Evidence for Its Incidence, Nature, and Psychological Impact From Three Daily Diary Studies. *Journal of Social Issues*, 57(1): 31–53.

Swim, J. K.; Mallett, R.; Russo-Devosa, Y.; and Stangor, C. 2005. Judgments of Sexism: A Comparison of the Subtlety of Sexism Measures and Sources of Variability in Judgments of Sexism. *Psychology of Women Quarterly*, 29(4): 406–411.

Appendix A. Meme Example

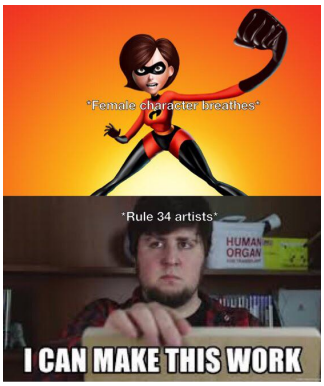

Figure 3: Meme example. All three female annotators classified it as *Objectification*, while no male annotator did.