SAMYRAD 2026

SMART ALGORITHMS & MODELS: YOUR REACH TO ADVANCED DIALOGUES · 5–6 OCTOBER 2026 · SEVILLE, SPAIN

SAMYRAD OPEN-ACCESS PROCEEDINGS | PUBLISHER: INTELLIGENT NUDGE

PAPER • OPEN ACCESS

An XAI Pipeline for Gender Bias Detection in Advertising Images

Fernando Martínez-Gil[1], Consolación Gil[1], Ainhoa Sánchez[1], Julio Gómez[1], Marjan Alavi [2]

[1] University of Almería, CIAIMBITAL, Almería, Spain

[2] McMaster University, Hamilton, Canada

Article DOI: 10.68286/samyrad2026.380

ABSTRACT

Gender bias remains pervasive in advertising, but large-scale assessment is challenging because manual coding is labour-intensive and automated methods may lack transparency and expert validation. This study presents an Explainable AI (XAI)-based pipeline to validate the Gender Bias Density Index (IDS) and examine its pedagogical potential. A corpus of 2,295 fashion and luxury advertising images was automatically coded using the locally deployed gemma4:26b model and evaluated against expert coding through agreement metrics, gradient boosting, permutation importance, and SHAP (SHapley Additive exPlanations). Results showed robust agreement for visually explicit variables but lower reliability for context-dependent features such as body agency and facial expression. Overall IDS agreement was moderate (ICC = 0.58), while the regression model achieved R^2 = 0.96 on internal test data and R^2 = 0.86 against expert coding. SHAP analysis showed that the most discriminative features broadly align with the theoretical structure of the IDS. The pipeline therefore supports scalable, interpretable bias screening for industry and provides a reproducible case study for teaching critical AI literacy in higher education.

KEYWORDS Explainable AI, advertising, gender bias, higher education, agency, facial expression

How to cite: Martínez-Gil, F., Gil, C., Sánchez, A., Gómez, J., & Alavi, M. (2026). An XAI pipeline for gender bias detection in advertising images. SAMYRAD 2026 Open-Access Proceedings. Intelligent Nudge. https://doi.org/10.68286/samyrad2026.380

Introduction

Gender bias in advertising has been studied for decades, with strong evidence that advertising media reproduce and reinforce stereotypes through physical attributes, social roles and visual codes. Structured instruments such as the Gender Bias Density Index (IDS) [1] have been developed to operationalise these patterns systematically. Gender asymmetries have also been documented in competence-related representations [2] and in the visual prominence of women and men in online images [3].

However, analysing gender bias at scale remains challenging. Manual content analysis is time-consuming and depends partly on the interpretive criteria of individual coders, while automated approaches require validation to determine whether they capture the intended social constructs reliably. This creates a need for methods that combine large-scale automatic analysis with expert validation and model interpretability.

Explainable Artificial Intelligence (XAI) provides a suitable framework for this purpose. Unlike conventional predictive models, which produce a score without explaining how it is obtained, tools such as SHAP make it possible to identify which specific visual variables increase or reduce the detected bias in each image. This transparency is valuable both for validating the measurement instrument and for understanding how automated systems operationalise socially meaningful variables.

This paper presents a validation pipeline for the IDS based on machine learning and XAI, implemented with open-source models running locally to ensure data privacy and reproducibility. The study is guided by three research

questions and a transversal pedagogical objective. RQ1: To what extent can the IDS variables be reliably coded by a locally deployed generative vision model, and which variables require human interpretive judgement? RQ2: To what extent does the IDS capture patterns that remain consistent across coding sources, so that a model trained on AI-coded data can predict the index computed from expert coding? RQ3: Which visual features empirically drive the IDS in the corpus, and how do they compare with the theoretical weights defined in the coding manual? As a transversal objective, the pipeline is structured so that each stage can be reused as a teaching unit for critical AI literacy in higher education. It provides a complete case study in which students can observe how a socially grounded construct, gender bias in advertising, is translated into operational variables, quantified through a structured index, evaluated through human-AI agreement, and finally interpreted using explainable machine learning techniques.

The paper is organised as follows: Section II reviews related work on advertising bias analysis and explainable AI; Section III describes the pipeline methodology; Section IV presents and discusses the results; and Section V sets out the conclusions and future work.

Background and related work

Gender bias can be defined as the tendency, conscious or unconscious, to favour one gender over another, producing inequalities across social, professional, educational and technological domains [2]. In advertising and digital content, its presence is well documented: men are associated with roles of power or action while women are given passive or decorative roles, systematic biases appear in the type of product advertised, and when both genders share the frame, the male figure tends to occupy the visual focus [3].

Automatic detection of these biases has developed mainly for text. Natural language processing techniques and supervised classifiers have proved effective in identifying biased language in job advertisements [4] or detecting stereotypes in news text [5]. Comparing lexicon-based approaches with transformer models shows that supervised learning, and BERT (Bidirectional Encoder Representations from Transformers) in particular, produces more accurate results, though it requires larger training datasets [6]. Automatic image analysis has moved more slowly, partly because of the greater technical complexity of computer vision and partly because visual bias patterns are harder to operationalise than textual ones. A further difficulty is the opacity of conventional predictive models, which produce a score without explaining how they reach it, limiting their usefulness as tools for auditing or validating measurement instruments.

Recent research on generative AI has made it possible to characterise more precisely which types of visual bias are reproduced automatically: European white phenotypes as the default, slim and athletic body types, straight hair as the female beauty standard, underrepresentation of older people especially women, systematic absence of visible disability, and gender asymmetries in facial expressions [7]–[9]. These dimensions match the variables captured by the IDS, ethnicity, body type, hair type, apparent age, disability, facial expression, which provides a solid empirical grounding for the instrument.

However, although recent studies of image-generative AI confirm visual representational biases, they do not establish whether a structured visual-bias index can be reliably automated or interpreted at the feature level. XAI research demonstrates the value of such feature-level explanations in other domains [10]. The present study bridges these strands by combining expert validation, predictive modelling and XAI within a single IDS-based framework.

Methods or analytical approach

This study adopts a quantitative, multi-stage validation design combining human–AI agreement assessment, predictive modelling and post-hoc explainability. This design was selected because validating a composite index requires evidence at three complementary levels, each corresponding to one research question: coding reliability against expert judgement (RQ1), internal and external predictive consistency (RQ2), and feature-level interpretability against the theoretical weighting (RQ3). A descriptive analysis of the IDS distribution across the corpus provides context for all three.

Dataset and inclusion criteria

The images used in this study come from the social media posts of internationally recognised fashion and luxury brands. The initial dataset contained 4,355 images, which were filtered using the following inclusion criteria: colour image, clear presence of one or two people with an apparent age of 18 or older, unambiguous gender identification, and representation of real people (excluding cartoons, caricatures or unidentifiable figures). After this filtering process, the dataset was fixed at 2,295 valid images. The distribution by gender composition is as follows: 1,459 images with one woman, 637 with one man, 89 with two women and 97 with two men. A further 13 images contained two people of whom only one was clearly visible and gender-identifiable; these were coded as single-person images using the formula for the visible person's gender. This distribution reflects the predominance of female figures in the advertising corpus, consistent with what has been documented in the literature [9].

The Gender Bias Density Index (IDS)

The Spanish acronym IDS (Índice de Densidad de Sesgo) is retained because it is the name used in the coding manual. The IDS is a measurement instrument developed by the research group of the project in [1] to quantify the level of gender bias in advertising images. The index brings together binary variables organised into four dimensions: biological biases (body type, hair type and hair color, skin conditions, disability representation), demographic biases (apparent age group, ethnicity), cultural biases (body agency, facial expression, visual focus, body exposure, spatial hierarchy) and socioeconomic biases (activity type).

Each variable is binarised as presence (1) or absence (0) of bias and is assigned a weight that differs according to the gender of the person depicted. The index takes the form of a normalised weighted sum, producing a continuous value in the range [0, 1], where 0 means no bias and 1 means maximum bias. Since the composition of people in the image determines both the applicable variables and the corresponding weights, the IDS is operationalised in four

specific formulas: image with one woman, one man, two women, or two men. For brevity, only the formulas for images with one woman (1) and one man (2) are shown here. The full specification, including the remaining formulas, the coded variables and their parameterisation, is available at [1].

Image with one woman:

$$IDS_i^F = \frac{\sum_{j \in J_F} x_{ij}^F w_j^F}{\sum_{j \in J_F} w_j^F} \qquad (1)$$

Image with one man:

$$\mathbf{IDS_i^M} = \frac{\sum_{j \in J_M} \mathbf{x_{ij}^M}\, \mathbf{w_j^M}}{\sum_{j \in J_M} \mathbf{w_j^M}} \qquad (2)$$

where:

- x_{ij}^F or x_{ij}^M = 1 if the image shows bias j, 0 otherwise
- w_{ij}^F or w_j^M = weight of bias j for female or male

Automatic feature extraction using local generative AI

The 2,295 images were coded using the generative AI model gemma4:26b, running locally on the University of Almería's own infrastructure. Using local models removes dependence on external paid services and ensures data privacy, which matters when the dataset includes images of real people.

Two servers were used. The first is a testing server, used for preliminary prompt-validation tests: a VPS with 8 vCPUs and 16 GB of RAM equipped with an NVIDIA RTX 2000E Ada Generation GPU (16 GB). The second is a production server, used for large-scale processing: it is equipped with two Intel® Xeon® Gold 5220 processors, 256 GB of RAM, and two GPUs, which serve the models. Both servers run an Ollama server that receives requests remotely from the application; the results reported in this article were obtained on the production server.

The prompt designed for feature extraction instructs the model to apply the IDS coding manual protocol, returning for each image and each identified person the categorical values of each variable and the corresponding binary value; the IDS is then calculated by applying the weighted formula described in Section III.B. The prompt includes explicit coherence rules that link categorical values to binary values, for example, if hair type takes the value “straight”, the corresponding binary variable must be 1, reducing the likelihood of internal contradictions in the model’s output.

Automatic coding validation: intercoder agreement

To assess the reliability of the automatic coding, a subsample of 85 individual images, 55 female and 30 male, was selected and manually coded by two experts. The experts first coded the images independently, following the IDS coding manual and without discussing their decisions. This independent round made it possible to identify variables with higher and lower levels of interpretive agreement. Discrepancies were then reviewed jointly and resolved by consensus, producing the final human expert coding used as the reference criterion for comparison with the AI output. Images with two people were excluded from this analysis because assigning biases to each subject separately introduces additional interpretive complexity.

Agreement between the automatic (AI) coding and the human coding was measured using Cohen’s Kappa [11], [12], calculated per variable and per gender on the binary presence/absence-of-bias values. The Kappa formula is:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad (3)$$

where P_o is the observed proportion of agreement and P_e is the proportion expected by chance, based on the marginal frequencies of each coder. The denominator ($1 - P_e$) normalises the index so that $\kappa = 0$ means chance-level agreement and $\kappa = 1$ means perfect agreement.

Interpreting Kappa requires looking at the value together with P_o and the confusion matrix, because a low Kappa can result from methodologically different causes. The analysis identified four qualitative profiles:

- Profile 1. Robust agreement ($\kappa \geq 0.50$): genuine agreement backed by sufficient variation in the data. Indicates that the variable can be reliably automated.
- Profile 2. Corpus homogeneity with bias: low Kappa with high P_o and P_e because the distribution of values is so homogeneous that chance agreement is already high. This does not indicate AI failure; it means that the bias is systematically present in nearly all images in the corpus.
- Profile 3. Real AI failure: low Kappa with moderate P_e and systematic false negatives in the confusion matrix. The AI does not detect the bias that the expert identifies, usually because the variable requires interpreting the narrative context of the advertisement.
- Profile 4. Corpus homogeneity without bias: all values are 0 for both coders. The type of bias captured by that variable does not appear in this corpus.

The full classification of variables by profile and gender is presented in Section IV.B and in Fig. 2 and Fig. 3

Machine learning model and IDS validation

The empirical validation of the IDS was designed as a two-phase process. In the first phase, a regression model based on gradient boosting (HistGradientBoostingRegressor from scikit-learn) was trained using as input variables the categorical values of the 12 IDS variables returned by the AI for each person, together with the number of people and the gender of each person, and the IDS calculated from that same coding as the target variable. Nominal variables were one-hot encoded (one binary column per category), apparent age group was ordinally encoded and disability representation was encoded as a single binary column; missing values were imputed with the most frequent category. To ensure independence of the external validation, the 99 images reserved for human coding (the 85 individual images of

Section III.D plus 14 two-person images) were excluded from this phase, leaving a modelling set of 2,196 images. This set was split 80% for training and 20% for testing, with strict separation of the two sets before any transformation to prevent data leakage. Model performance was measured using the coefficient of determination (R^2) and mean absolute error (MAE). The model hyperparameters were: L2 regularisation: 10.0, learning rate: 0.2, maximum iterations: 300, maximum leaf nodes: 15, minimum samples per leaf: 20.

In the second phase, the trained model was applied without retraining to the 99 images coded manually by human experts (the gold set), including the two-person images, since the IDS formula applies equally to those cases. The expert-coded categorical values were transformed with the encoders fitted on the AI-coded data, so that categories absent from training (e.g. the experts' "other" or "not applicable" codes) yield all-zero indicator columns. The target variable was the IDS calculated from the human coding [1]. This phase constitutes the external validation of the index: a high R^2 against the expert criterion shows that the IDS captures objectively detectable patterns that go beyond the coding source, and the gap between the internal (Phase 1) and external (Phase 2) R^2 reflects how far the expert coding departs from the AI coding, linking model validation to the agreement results of Section III.D.

Two complementary techniques were used for model interpretability: permutation importance, which quantifies each variable's contribution to the prediction by measuring the drop in R^2 when its values are randomly shuffled in the test set; and SHAP, which, unlike permutation importance, allows us to examine both the average importance of each variable and the direction of its effect — whether the presence of a category (indicator value = 1) tends to increase or decrease the predicted IDS. This dual capacity, magnitude and direction, makes it possible to compare the empirical weights learned by the model with the theoretical weights in the IDS manual, and to assess how far the model's behaviour matches the theoretical design of the instrument.

Pedagogical use of the validation pipeline

From a pedagogical perspective, the pipeline was designed as a modular workflow transferable to teaching in artificial intelligence, digital literacy and social bias analysis. Each stage can be used as an instructional unit: the coding manual introduces the operationalisation of abstract social constructs; the automatic coding stage illustrates the possibilities and limits of generative AI for image interpretation; the reliability analysis shows why algorithmic outputs must be validated against human judgement; and the XAI stage lets students examine which variables drive the predictions and whether they match the theoretical design of the instrument.

The pipeline thus supports active learning activities in which students compare human and AI codings, interpret confusion matrices and discuss whether model errors reflect technical limitations, corpus homogeneity or genuinely ambiguous social meanings, developing critical AI literacy through a concrete and reproducible bias detection task.

Results and Discussion

Distribution of the IDS in the corpus

The IDS calculated automatically for the 2,295 images shows a mean of 0.443 (SD = 0.114) and a range from 0.0 to 0.92. The histogram (Fig. 1) reveals an irregular shape with two frequency peaks around 0.45 and 0.55, and a progressive decline towards the extremes. Although the IDS is expressed as a continuous value in the range [0, 1], it is calculated from a weighted sum of binary variables. Therefore, local accumulations around specific values are expected, since images with similar combinations of activated bias indicators may obtain similar IDS scores. The low proportion of values below 0.2 or above 0.75 indicates that the advertising corpus tends towards moderate levels of bias, with neither neutral nor strongly biased images predominating.

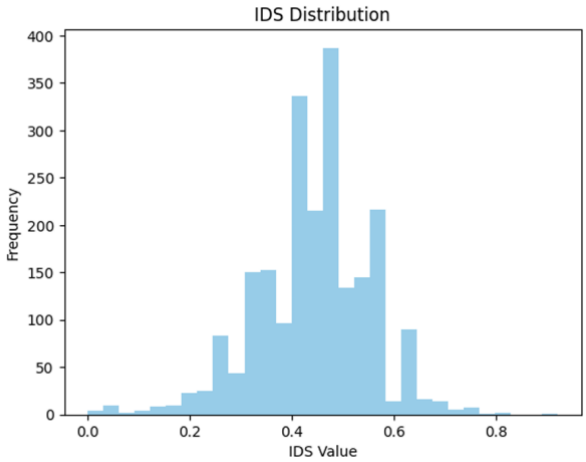


Fig. 1. Histogram showing the distribution of the Gender Bias Density Index (IDS).

Images at the extremes of the index illustrate the discriminative capacity of the IDS. The image with the highest detected bias (IDS = 0.92) is an Adidas post featuring a young woman with a European white phenotype, slim build, straight blonde hair, blemish-free skin, an unstable or forced body posture, tight-fitting clothing, a camera focus on the body and a self-care role: a combination of indicators that activates nearly all the biases in the female IDS formula. At the other end, the image with IDS = 0.0 belongs to the Hermès brand and shows two women, one with an Afro-descendant phenotype, natural curly hair and a serious expression, and another whose characterisation is not applicable due to low visibility, resulting in no detectable biases under the coding manual.

Intercoder agreement and IDS concordance: reliability of automatic coding (RQ1)

The Kappa analysis results, shown in Fig. 2, give a mean of κ = 0.397 for female images (N = 55) and κ = 0.529 for male images (N = 30), excluding trivial variables. This moderate overall value conceals, however, meaningful qualitative differences between variables that the profile analysis can unpack.

Profile 1: Robust agreement. The variables that exceed the $\kappa \geq 0.50$ threshold are: ethnicity (κ = 0.855, Po = 92.7%, Pe = 50.0% for female; κ = 0.864, Po = 93.3%, Pe = 50.9% for male), body exposure (κ = 0.595, Po = 80.0%, Pe = 50.6% for female), hair color (κ = 0.614, Po = 90.9%, Pe = 76.4% for female; κ = 0.516, Po = 90.0%, Pe = 79.3% for male) and activity type for female (κ = 1.000, Po = 100%, Pe = 93.0%). For male images, hair type (κ = 0.706, Po = 86.7%, Pe = 54.7%), disability (κ = 1.000, Po = 100%, Pe = 87.6%) and visual focus (κ = 1.000, Po = 100%, Pe = 93.6%) are also in this profile. In all these

variables Po clearly exceeds Pe, confirming that the agreement is genuine and not an artefact of the corpus distribution. These are visually and objectively detectable features: ethnicity, body framing or activity type present unambiguous cues for the vision model. The contrast between hair type for male images (P1, κ = 0.706, Po = 86.7%, Pe = 54.7%) and female images (P3, κ = 0.350, Po = 65.5%, Pe = 46.9%) shows how the same variable can have very different reliability depending on gender, possibly because the distinction between straight and wavy hair is more ambiguous in the female fashion and luxury corpus. Fig. 2 shows this profile in green and Fig. 3(a) shows the confusion matrix for ethnicity.

Profile 2: Corpus homogeneity with bias. The variables skin conditions (κ = 0.151, Po = 85.5%, Pe = 82.9% for female; κ = 0.286, Po = 76.7%, Pe = 67.3% for male), disability for female (κ = 0.486, Po = 96.4%, Pe = 92.9%), apparent age (κ = 0.347, Po = 85.5%, Pe = 77.7% for female; κ = 0.474, Po = 93.3%, Pe = 87.3% for male), visual focus for female (κ = −0.031, Po = 89.1%, Pe = 89.4%), body agency for male (κ = −0.035, Po = 93.3%, Pe = 93.6%) and activity type for male (κ = 0.000, Po = 96.7%, Pe = 96.7%) share the defining feature of Profile 2: Pe is close to or equal to Po, so Kappa is not measuring AI failure but corpus homogeneity. The luxury and fashion advertising analysed systematically shows perfect skin, absence of disability, young or middle-aged subjects and men in direct product-showcase roles, meaning these variables barely distinguish between images. This result is itself a finding: the biases in Profile 2 cannot be quantified by Kappa precisely because they are always present in the corpus. Fig. 2 shows this profile in orange and Fig. 3(b) shows the confusion matrix for skin conditions.

Profile 3: Real AI failure. The variables body type for female (κ = 0.189, Po = 67.3%, Pe = 59.7%), hair type for female (κ = 0.350, Po = 65.5%, Pe = 46.9%), body agency for female (κ = −0.012, Po = 72.7%, Pe = 73.1%) and facial expression (κ = 0.217, Po = 80.0%, Pe = 74.5% for female; κ = 0.483, Po = 73.3%, Pe = 48.4% for male) differ from Profile 2 in that Pe is moderate and Po does not exceed it significantly, revealing a real disagreement and not a homogeneity effect. The confusion matrices show systematic false negatives: the AI does not detect the bias that the expert identifies. The most extreme case is body agency for female (κ ≈ 0, Po = 72.7%, Pe = 73.1%), where the AI performs no better than chance at distinguishing a decorative female posture from a functional one — a distinction that requires reading the full narrative context of the advertisement. The combined weight of these variables in the female IDS formula (body agency = 4, facial expression = 4, hair type = 3, body type = 3) accounts for 28% of the index, with a direct impact on external validation (Section IV.C). The gender asymmetry in facial expression, better detected for male images (κ = 0.483, Po = 73.3%, Pe = 48.4%) than for female (κ = 0.217, Po = 80.0%, Pe = 74.5%), reflects the fact that male biased expressions (serious/authoritative) are more objectively observable than female ones (seductive/vulnerable), which carry a heavier cultural load. For male images, with Po exceeding Pe by 25 points, it falls just below the Profile 1 threshold and reflects moderate rather than failed agreement. Fig. 2 shows this profile in red and Fig. 3(c) shows the confusion matrix for body agency.

Profile 4: Corpus homogeneity without bias. Body type for male (trivial κ, Po = 100%, Pe = 100%; all values are zero because the instrument does not code male slimness as a bias indicator) and body exposure for male (trivial κ, Po = 100%, Pe = 100%; body exposure is absent in male luxury advertising) show trivial agreement because the bias simply does not appear in this corpus. The fact that body objectification and slimness as an aesthetic ideal are biases that are absent for men but present for women is itself a finding about the structural asymmetry of gender bias in the advertising analysed. Fig. 2 shows this profile in blue and Fig. 3(d) shows the confusion matrix for body type.

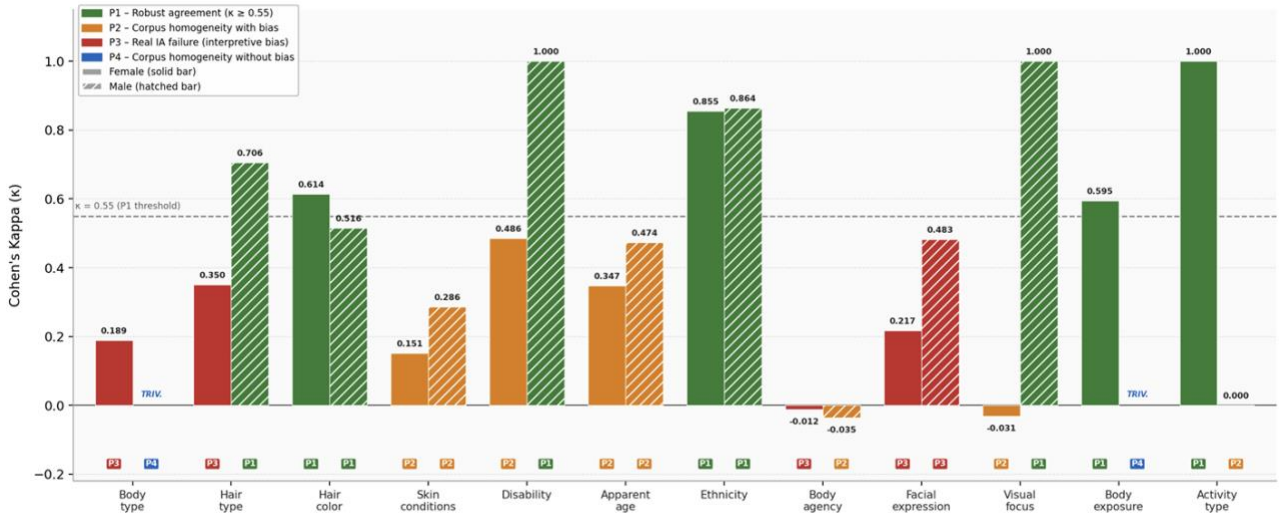


Fig. 2. Cohen's Kappa intercoder agreement between AI and human coding, by variable and gender (female N = 55, male N = 30).

Pedagogically, the four-profile interpretation shows students that a low Kappa does not always mean model failure: it may reflect corpus homogeneity or the model's inability to read culturally loaded cues, as the confusion matrices in Fig. 3 make concrete by distinguishing which biases are visually explicit, which are inferred from context, and which the model misses.

Variable-by-variable Kappa analysis gives only a partial picture of reliability, as it evaluates individual binary components that are sensitive to class imbalance in the corpus. To assess whether the automatic coding produces an overall IDS equivalent to the expert's, three complementary analyses were carried out on the N = 85 individual images (N = 55 female, N = 30 male).

The Intraclass Correlation Coefficient (ICC), a standard psychometric tool for validating composite scales [13], gave ICC(A,1) = 0.582 (95% CI = [0.42, 0.71], p < 0.001) overall; 0.489 ([0.26, 0.67]) for female images; and 0.593 ([0.25, 0.79]) for male images. According to Koo and Li's criteria [13] (below 0.50 = poor, 0.50–0.75 = moderate, above 0.75 = good), reliability is moderate overall and for male images, and at the boundary between poor and moderate for female images, where the Profile 3 variables concentrate; this is acceptable for an instrument that includes partly subjective interpretive variables. The closeness of ICC(A,1) = 0.582 and ICC(C,1) = 0.587 confirms that the systematic bias between the two coding sources is minimal, so the instrument can be applied automatically, with acceptable limitations, for comparative analyses at scale.

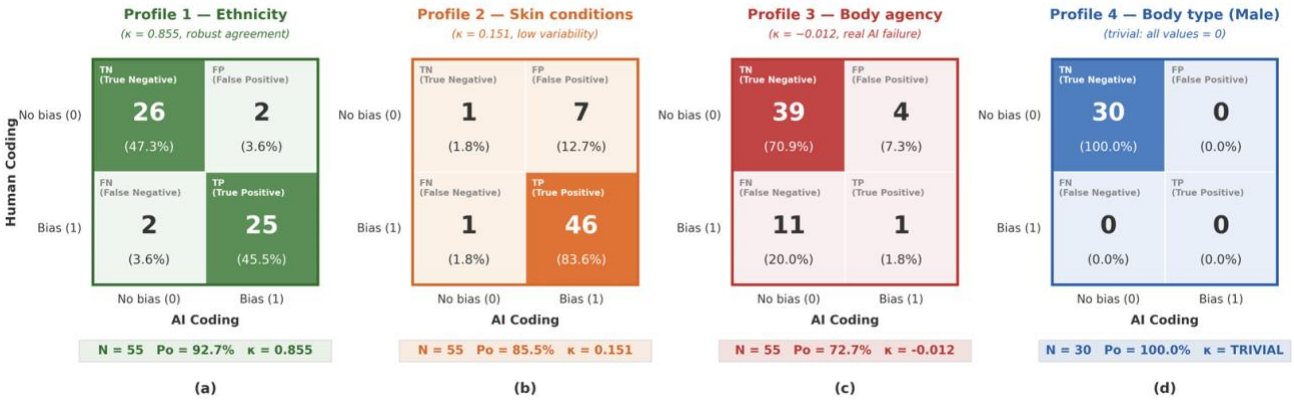


Fig. 3. Representative confusion matrices for the four agreement profiles between AI and human coding (absolute values and percentages).

To evaluate the practical image-by-image error, Bland-Altman analysis was used [14]. The overall mean bias is +0.020 (95% CI = [−0.003, +0.043]), equivalent to just 2 points out of 100, with no practical relevance. The limits of agreement (LoA = [−0.191, +0.231]) indicate that empirically, 92.9% of the images fell within these limits between −19 and +23 points out of 100. A statistically significant proportional error was found for female images (r = −0.376, p = 0.005): the AI tends to overestimate the IDS for low-bias images and underestimate it for high-bias ones, a pattern consistent with the greater weight of Profile 3 variables in images with higher IDS values. For male images the bias is slightly larger (+0.039) but constant (p = 0.161). Decomposed by variable, it originates mainly in skin conditions and facial expression, where the AI codes bias more often than the expert (93% vs 70% and 73% vs 47% of male images, respectively), whereas hair color and apparent age, whose binarisation rules differ between genders, contribute negligibly.

Spearman correlation was used for image ranking because it is more robust to the non-normal distribution of the IDS and to the presence of outliers. The results (ρ = 0.623 overall; ρ = 0.480 for female; ρ = 0.649 for male; p < 0.001 in all cases) indicate moderate ordinal concordance. The AI correctly identifies which images in a campaign carry the highest bias load in most cases, which is the most practically relevant use of the instrument for communication teams.

The analysis of systematic error (IDS_AI − IDS_Expert) by brand reveals a methodologically significant pattern. Luxury brands with a stronger narrative component show negative errors, the AI underestimates the IDS, possibly because the bias is expressed through staging and mise-en-scène codes that go beyond the vision model's capacity: Dior (error = −0.046), Chanel (−0.039). By contrast, sports and accessible brands show positive errors where the bias attributes are more directly observable: Adidas (+0.056), Nike (+0.042). The most extreme case is Rolex (+0.109), where the AI overestimates the IDS in both female (+0.137) and male (+0.083) images: this brand favours mature subjects, whose visible signs of age the expert codes as non-idealised attributes, whereas the AI tends to code them as unblemished skin, youthful appearance and serious expression. This pattern suggests that the reliability of automatic coding is not uniform across the corpus but varies systematically with the advertising genre of the brand, which is a specific limitation of the instrument that should be taken into account when interpreting results at the level of individual campaigns.

Empirical validation of the IDS using gradient boosting (RQ2)

Phase 1: Internal consistency

The HistGradientBoostingRegressor trained on the AI-coded binary features from the 2,196 images achieved R^2 = 0.99 on the training set (1,757 images, 80%) and R^2 = 0.96 on the test set (439 images, 20%), with MAE = 0.004 and MAE = 0.010 respectively. The three-percentage-point difference between training and test is expected and does not indicate overfitting, but rather solid generalisation to unseen data. The test MAE of 0.010 corresponds to about one percentage point on the IDS scale. This shows that the model reproduces the logic of the IDS without accessing the formula, validating the internal consistency of the instrument.

Phase 2: External validation against human criterion

When the trained model was applied to the 99 manually coded images, R^2 = 0.86 and MAE = 0.031, equivalent to a mean error of 3.1 percentage points on the index scale. The 10-percentage-point drop from the Phase 1 test R^2 (0.96) reflects combinations of categories that are rare or absent in the AI-coded training data: the experts make systematic use of the "other" and "not applicable" codes (e.g. in body agency, facial expression, body exposure and activity type), which the AI seldom returns, and binarise some variables differently (male body type is never coded as bias; serious male expressions less often). These are the discrepancies

documented by the Kappa profiles in Section IV.B, so the external R^2 is consistent with the moderate agreement between coding sources (ICC(A,1) = 0.58) while confirming that the model has learned the logic of the index rather than the idiosyncrasies of the AI coding.

In particular, body agency (weight = 4) and facial expression (weight = 4), the Profile 3 variables with the largest AI–human disagreements, make up 16% of the female index, so their coding differences have a proportional impact on the IDS; this link between the reliability analysis and the model validation is one of the contributions of this work.

Variable importance: coherence with the theoretical framework (RQ3)

The permutation importance analysis on the test set identified five variables as the most influential in IDS prediction: gender_b (1.107), body_exposure_a (0.277), ethnicity_a (0.273), gender_a (0.197) and hair_type_a (0.101). The dominant role of gender_b is methodologically expected: this variable encodes whether the image contains one or two people, determining which IDS formula to apply. Permuting its value forces the model to use the wrong formula, causing the largest drop in performance (a value above 1 indicates that the permuted R^2 becomes negative). This confirms that the model has learned the compositional structure of the index rather than simply individual variable patterns.

Among the substantive variables, body_exposure_a ranks second; its most influential category groups forms of body exposure other than tight clothing or partial nudity, reflecting that body exposure type is highly discriminative between images with different bias levels. Ethnicity ranks third, consistent with its theoretical weight (weight = 4 for women, 3 for men) and the high prevalence of European white phenotypes in the corpus. Gender_a appears in fourth place and, like gender_b, determines which IDS formula to apply. Fifth place goes to hair_type_a despite having a lower theoretical weight (weight = 3 for women, 2 for men), suggesting that its high frequency in the corpus amplifies its discriminative capacity.

The SHAP analysis (Fig. 4) confirms and nuances these results. In the global beeswarm plot, red dots indicate images where the category is present (indicator value = 1): when they are concentrated to the right of the axis, the presence of that category increases the predicted IDS. The categories ethnicity_a_white_european, hair_type_a_straight and body_type_a_thin show this pattern clearly and consistently in the global analysis. By contrast, the red dots of body_exposure_a_other are concentrated to the left of the axis: this category groups forms of body exposure other than tight clothing or partial nudity, i.e. the non-biased value of the variable, so its presence lowers the predicted IDS. The model has thus learned the direction of each category, in agreement with the design of the instrument.

It is worth noting that variables with high theoretical weight but low variability in the corpus, such as disability and skin conditions, classified in Profile 2 in the Kappa analysis (Section IV.B), are absent from the top of both analyses. This is not a flaw in the instrument but a direct consequence of corpus homogeneity: when a variable takes the value 1 in more than 90% of images, it contributes to the IDS in a constant way but has no capacity to discriminate between images. Taken together, the permutation importance and SHAP analyses suggest that part of the theoretical weighting in the IDS reflects biases that are systematically present throughout the luxury and fashion advertising corpus, rather than features that differentiate more from less biased images within it.

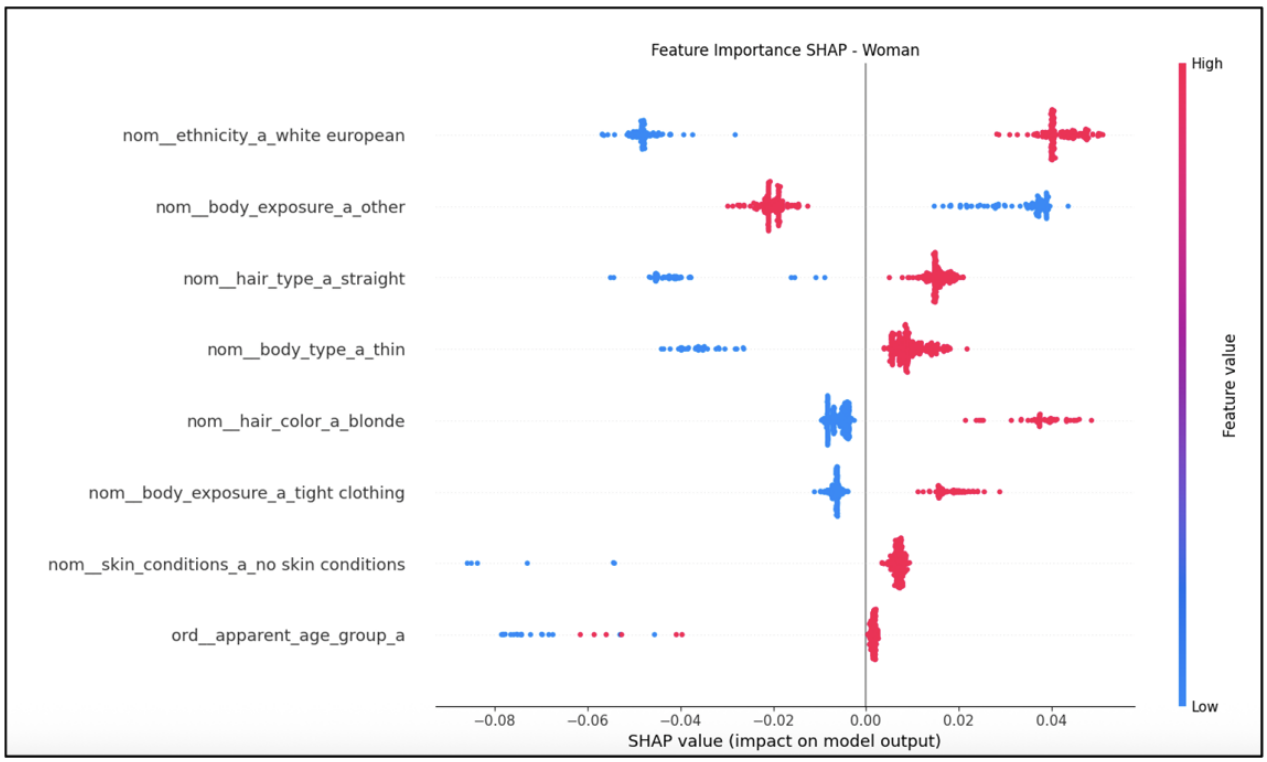


Fig. 4. SHAP analysis of the main features driving IDS predictions (global beeswarm plot).

For teaching purposes, SHAP lets students contrast the manual's theoretical weights with the empirical importances and discuss why high-weight variables can lack discriminative power in a homogeneous corpus.

Practical implications

For industry, the results have three practical implications. Communication and brand teams could use the pipeline as a pre-publication screening tool, as the AI correctly ranks the most biased images in a campaign in most cases (ρ = 0.62). Profile 1 variables can be reliably automated, whereas Profile 3 variables, such as body agency and facial expression, should be flagged for human review. Finally, the brand-level error pattern suggests that screening thresholds may require calibration by advertising genre, since narrative luxury brands tended to be under-scored, whereas sports brands tended to be over-scored in this corpus.

Conclusion

This work has shown that empirically validating a gender bias index using machine learning and XAI is feasible, with models running locally to ensure privacy and reproducibility.

In answer to the research questions, the results show that (RQ1) visually explicit variables such as ethnicity, hair color and body exposure can be reliably automated, whereas context-dependent variables such as body agency and facial expression still require human judgement; (RQ2) the IDS captures patterns consistent across coding sources, with a model trained on AI-coded data predicting the expert-coded index with R^2 = 0.86; and (RQ3) the features that empirically drive the index partially coincide with those of highest theoretical weight, the main discrepancies arising from variables that are near-universal in the corpus and therefore lack discriminative power.

The intercoder analysis reveals four agreement profiles with different methodological implications. Variables with robust agreement (Profile 1: ethnicity, hair color, body exposure and, for men, hair type) can be directly automated. Profile 2 variables (skin conditions, disability and apparent age) show low Kappa not because the AI fails but because the bias is so systematically present in the luxury advertising analysed that it does not discriminate between images: the low Kappa itself is a finding about the nature of the corpus. Profile 3 variables (body agency and facial expression for female) expose a real limit of the vision model in interpreting the narrative context of the advertisement, and their combined weight in the female IDS formula (16%) is consistent with the drop in R^2 in the external validation (R^2 = 0.86).

The SHAP analysis confirms that the variables with the greatest discriminative power broadly match those with the highest theoretical weights in the instrument, with the exception of Profile 2 variables, whose theoretical importance does not translate into empirical importance within this corpus. The brand-level error analysis further shows that the reliability of automatic coding varies systematically between narrative luxury brands and sports brands, which is a specific limitation to bear in mind when applying the instrument at the level of individual campaigns.

From an educational perspective, the main contribution of the work is the design of a replicable XAI-based learning scenario for social bias detection. The pipeline allows students to engage with the full cycle of responsible AI analysis: defining a socially meaningful construct, operationalising it through variables and weights, comparing human and automatic coding, evaluating reliability, training predictive models and interpreting their decisions through XAI. This makes the IDS validation process a useful pedagogical resource for promoting critical AI literacy, particularly in contexts where students need to understand both the technical potential and the social limitations of automated decision-making systems.

Future work will therefore not only focus on improving the predictive pipeline, but also on adapting it into structured teaching activities for higher education. These activities may include guided coding exercises, human-AI comparison tasks, interpretation of Kappa profiles and confusion matrices, SHAP-based model explanation, and counterfactual redesign of biased images using DiCE (Diverse Counterfactual Explanations) to generate image-level recommendations, completing the pipeline from bias detection to bias correction. Such activities would make it possible to evaluate the pipeline not only as a validation method, but also as an educational intervention for teaching explainable and socially responsible AI.

Declarations

Funding: This work is supported by Project “21-ID25, Femvertising, AI and Body Diversity: Towards Inclusive and Transformative Advertising”, Instituto de la Mujer (Ministry of Equality, Spain).

AI use disclosure: AI was used for translation assistance and English language editing.

Acknowledgement

The authors thank Ivan González for his help collecting images from social media, the members of the HUM-1022 and TIC-221 UAL research groups for their collaboration in developing the IDS and their participation as expert coders, and the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this manuscript.

References

[1] IDS (Índice de Densidad de Sesgo). IDS Coding Manual for Gender Bias Detection. Project 21-ID25 (UAL) [Online]. Available: https://www.ualtech.es/projects/gender_bias_detection/#top. [Accessed: Aug. 21, 2026]

[2] R. Poveda-Bautista, J. A. Diego-Mas, H. Gonzalez-Urango, and C. Corona-Sobrino, “Conscious and unconscious gender bias in competence evaluations: Mental representations of project managers,” Complexity, vol. 2025, 2025, doi: 10.1155/cplx/7974362.

[3] D. Guilbeault, S. Delecourt, T. Hull, B. S. Desikan, M. Chu, and E. Nadler, “Online images amplify gender bias,” Nature, vol. 626, no. 8001, pp. 1049–1055, 2024, doi: 10.1038/s41586-024-07068-x.

[4] R. Frissen, K. J. Adebayo, and R. Nanda, “A machine learning approach to recognize bias and discrimination in job advertisements,” AI & Society, 2022, doi: 10.1007/s00146-022 01574-0.

[5] C. H. M. Landa, Implementación de NLP y análisis temático para la detección de sesgos de género. Tlaquepaque, Jalisco, México: Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO), 2024.

[6] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao, "Detecting gender stereotypes: Lexicon vs. supervised learning methods," in Proc. 2020 CHI Conf. Human Factors in Computing Systems, pp. 1–11, 2020, doi: 10.1145/3313831.3376488.

[7] J. Shuford, "Examining ethical aspects of AI: Addressing bias and equity in the discipline," Journal of Artificial Intelligence General Science (JAIGS), vol. 3, no. 1, pp. 262–280, 2024, doi: 10.60087/jaigs.v3i1.119.

[8] L. Sun, M. Wei, Y. Sun, Y. J. Suh, L. Shen, and S. Yang, "Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI," Journal of Computer Mediated Communication, vol. 29, no. 1, Art. no. zmad045, 2024, doi: 10.1093/jcmc/zmad045.

[9] S. L. Grau and Y. C. Zotos, "Gender stereotypes in advertising: A review of current research," International Journal of Advertising, vol. 35, pp. 761–770, 2016, doi: 10.1080/02650487.2016.1203556.

[10] S. Nejadshamsi, C. H. Chu, K. McGilton, X. Li, C. E. Ronquillo, and S. Abbasgholizadeh-Rahimi, "Evaluation and improvement of algorithmic fairness for COVID-19 severity classification using explainable artificial intelligence-based bias mitigation," JAMIA Open, vol. 9, no. 1, Art. no. ooaf171, 2026, doi: 10.1093/jamiaopen/ooaf171.

[11] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37 46, 1960, doi: 10.1177/001316446002000104.

[12] K. A. Neuendorf, The Content Analysis Guidebook, 2nd ed. Thousand Oaks, CA, USA: Sage, 2016.

[13] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," Journal of Chiropractic Medicine, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.

[14] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," The Lancet, vol. 327, no. 8476, pp. 307–310, 1986, doi: 10.1016/S0140 6736(86)90837-8.