Abstract

English

As face-analysis systems permeate daily life—from security checkpoints to social-media beauty filters—their opacity gives rise to pressing questions of fairness, accountability, and transparency. In response, this dissertation brings together insights from artificial cognition, feminist theories, and critical race studies to develop an interdisciplinary methodology, aiming to clarify algorithmic behavior and expose the broader socio-political implications of computational vision. The methodology is applied through a set of interrelated case studies spanning face verification, augmented-reality beauty filters, generative AI for synthetic faces, gender classification, and the use of race in machine-learning fairness frameworks.

From a technical perspective, the work introduces concept-based explainable AI methods tailored to face-verification systems. These methods combine targeted perturbations with semantic mappings of facial landmarks, then use large language models to translate the outputs into explanations that align with human cognitive processes. The resulting explanations are designed to be clear and relevant to people from different backgrounds. Their effectiveness is evaluated quantitatively, and they are implemented into interactive prototypes to ensure accessibility across demographic groups. In the context of augmented reality filters, the dissertation builds a transparency-driven "Disclaimer Block" that exposes transformation parameters and audits how gender classification is embedded in visual aesthetics. Using controlled generation pipelines, the research isolates attractiveness as a variable to investigate how it affects downstream gender classification accuracy, revealing systematic biases. Finally, it critiques the use of rigid racial taxonomies in machine-learning fairness frameworks, proposing context-sensitive, non-essentialist alternatives informed by mixed-race positionality and situated European legal norms.

Taken together, these empirical studies challenge the idea that technical fixes alone can ensure trustworthy AI, advocating instead for a reframing of how algorithmic systems are designed, explained, and governed. The result is a set of sociotechnical interventions that aim to render facial analysis systems more transparent, equitable, and critically aware of the visual politics they encode.