Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

KeepA(n)I: Social Stereotypes in and Social Norms for Computer Vision

Evgenia Christoforou[1,3], **Nicolas Nicolaou**[2], **Efstathios Stavrakis**[2], **Jahna Otterbacher** [1,3]

[1]CYENS - Centre of Excellence, Nicosia, Cyprus
[2]Algolysis Ltd, Nicosia, Cyprus
[3] Open University of Cyprus, Nicosia, Cyprus

Abstract

The KeepA(n)I platform facilitates the auditing of computer vision systems that tag images, which aid visual communication on the Web and social media, from content moderation to the development of new apps and tools. In particular, KeepA(n)I enables a broad set of stakeholders to scrutinize a process of interest that embeds an image tagger for issues of social stereotyping, while also examining the social norms that humans apply to the observed AI behaviors. KeepA(n)I's approach, and its use of the power of the crowd, can aid the stakeholders in receiving responses to both descriptive and normative questions (i.e., which stereotyping behaviors are observed and if they are considered problematic by a given "crowd" for an intended context). We provide an overview of the platform, its key features, and a discussion via a use case on the diverse set of stakeholders that can benefit from it.

Introduction

There has been a rapid democratization of AI tools and services, which can benefit stakeholders in Web and social media communication, including those who analyze communication, moderate content, or develop new processes and applications. However, AI can inadvertently perpetuate or amplify biases, leading to outcomes that may reinforce discrimination or unfair treatment (Ntoutsi et al. 2020; Ferrara 2024; Baldassarre et al. 2023). Thus, there is a need for constant monitoring (Barlas et al. 2022). In the near future, regular audits of AI systems are likely to become legal requirements, ensuring that they are used responsibly and ethically. However, in light of the rise of AI systems with more unpredictable,"human-like" behaviors, researchers have described a need for an ecosystem of AI evaluation techniques (Weidinger et al. 2023) to involve diverse stakeholders.

When auditing an AI system that exhibits "human-like" behaviors, one must consider how to measure the possible expression of social stereotypes (Steed and Caliskan 2021; Papakyriakopoulos et al. 2020). Stereotypes concern the beliefs that people hold about others, usually based on their social categorization (Judd and Park 1993) (e.g., gender, age, race, ethnicity). As social stereotypes are transmitted between people, they can also be propagated by AI (Barlas et al. 2021; Marinucci, Mazzuca, and Gangemi 2023). Whereas stereotypes concern beliefs about how people are or what they "should be," social norms are the beliefs or standards that people hold about *behaviors*, i.e., what is acceptable behavior and what is not. This includes what one should (not) say about others (Zlatev and Blomberg 2019). Thus, taken as a whole, the social norms of a community define its "socially acceptable way of living" (Kandori 1992). In the same way, AI systems deployed in a certain community must, through their output, align with the acceptable behavior(s) for that community (Otterbacher 2023).

KeepA(n)I examines the expression of stereotypes and how they are reflected in biases shared by groups of people interacting with the system (i.e., social norms). It is a human-in-the-loop approach, engaging people in the evaluation process, via paid or volunteer crowdsourcing. It facilitates diverse (e.g., across cultures) and dynamic (e.g., across contexts and time) evaluation of social norms.

The KeepA(n)I Platform

The KeepA(n)I platform: (1) identifies social stereotypes expressed by image taggers when analyzing people images; (2) determines social norms for the tagger's behaviors, using crowd input; (3) promotes fairness by supporting non-expert developers and researchers in auditing AI systems; (4) raises awareness of computer vision stereotypes through dynamic empirical evidence.

Currently, the platform is configured to support stakeholders in understanding the behaviors of AI image taggers before they embed them into a process or system of their interest (hereon: "System"), to create an enhanced "AI-enabled System." The platform offers a *user mode* and an *admin mode*, allowing administrators to review and approach the launch of a crowdsourcing campaign. In this demo, we focus on the *user mode*.[1]

Input and Output of the Opaque *AI Tagger*

We assume that the *System* feeds part of its input through the *AI tagger*, and that in turn, the output of the *AI tagger* affects the decisions made and the output of the resulting AI-enabled *System*. For the platform to audit the input and output of the *AI tagger* with regards to the *System* at hand, it introduces a set of functionalities as described below.

[1]An account can be created via: https://keepani.algolysis.com. A demo video can be found at: https://youtu.be/TUw6XfLfFkA

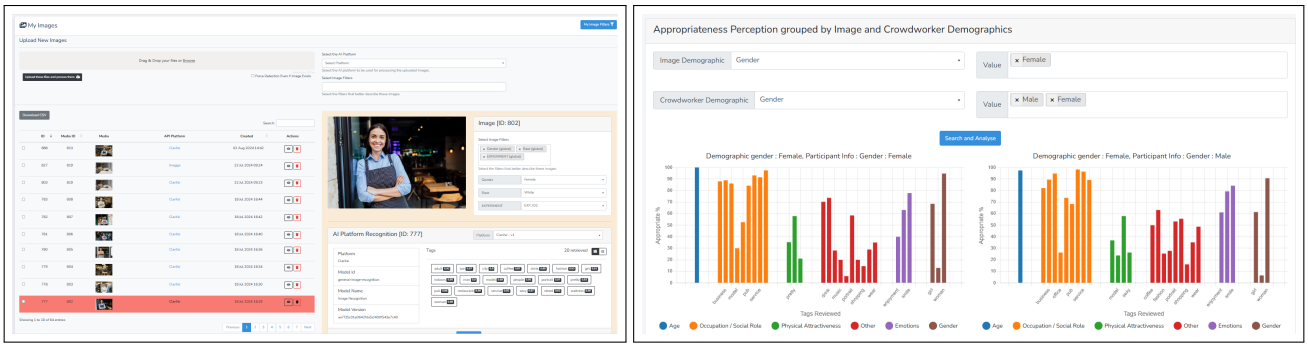

Figure 1: (Left) Image Uploading feature. (Right) Dynamic reporting on social norms after crowdsourcing evaluation.

Image Uploading: Through *My Images*, a user uploads an image to the platform (see Figure 1), and can label the image with values specified in *My Image Filters* .

Image Labeling: My Image Filters offers two key functions: defining image demographics (e.g., gender, race, age) and grouping images with labels for easier access and evaluation. The platform includes two global filters for Gender and Race and the option to create custom filters.

Image Processing: KeepA(n)I currently features two taggers: Clarifai (Clarifai 2025) and Imagga (Imagga 2025), calling their APIs and importing all generated tags into the platform (see Figure 1). Additionally, through the option *Custom Tags*, the platform allows the user to upload and tag the selected image with tags generated or collected through other means, which relate to their specific problem. Thus, KeepA(n)I allows not only the evaluation of *Systems* enhanced by Imagga and Clarifai, but also, other image taggers and annotation means (e.g., crowdsourcing).

Social Stereotype Identification

The *AI-Tagger Report* feature allows users to evaluate social stereotypes expressed via image descriptions generated by the selected image tagger. Users can filter images by criteria and tags, then request a report on potential stereotypes detected. The generated report provides statistics on the demographics of the selected images and the tag frequencies over all images. Heat-maps illustrate word pair frequencies, exposing potential biases in tagger behavior. This process can reveal patterns such as stereotypical tag pairings (e.g., "sexy" and "young" describing images of women but "serious" and "fine-looking" describing men). Additionally, the report details error rates in gender tagging (where ground truth is available) and tracks specific word usage across demographic combinations. This information can aid users in assessing the impact of tag usage on system development.

Social Norms for an AI-enabled System

Under the feature *My Experiments*, the user provides details about the *System*'s objectives and target audience. This is used to generate the use scenario under which social norms will be determined, as well as the crowdsourcing parameters. In particular, when selecting a crowd over a crowdsourcing platform, KeepA(n)I's protocol will also aim to collect the perception of users who match the demographics of the target users of the *System*, as described by the developer. The user also specifies the image tagger to be used by the *System*, together with a set of evaluation images.

Once a *New Experiment* request is placed by the user, the platform will generate one or more crowdsourcing campaigns to facilitate the identification of social norms to which the *System* should adhere. Following KeepA(n)I's crowdsourcing protocol, created campaigns will target specific demographic groups of the population, who will be asked to judge, through the crowdsourcing task presented to them, the tags' appropriateness when describing a specific image and *System* scenario of use. Furthermore, the KeepA(n)I platform will adjust the various crowdsourcing parameters, such as the task duration (i.e., how many images workers should evaluate in a single task) and the reliability of results (i.e., how many unique responses should be collected per image). Under *My Campaigns*, users can view the status of campaigns generated within an experiment.

Once all campaigns under an experiment are completed, a report is generated under *My Experiments*. The report provides an overview of the recruited crowdworkers' demographics (as collected by the crowdsourcing platform and a self-reporting questionnaire). An overview of their perceptions of which tags are (in)appropriate for a specific *System* scenario of use, per image gender and race, is provided. This information accompanies the crowdworkers' perception of the category the specific tag falls under, when perceived as appropriate or not, i.e., tags relating to race, gender, age, emotions, traits, occupation/social role, physical attractiveness, inflammatory and other (Barlas et al. 2019). Furthermore, the user is provided the flexibility to view the social norms (i.e., appropriateness perceptions) grouped by image and crowdworker demographics (see Figure 1).

Applications and Impact

KeepA(n)I is useful to developers wanting to audit an AI-enabled system's behavior in target cultural and geographical contexts. KeepA(n)I not only identifies possible stereotypes in the image dataset on which the system is being developed, but also, with the aid of crowdworkers, determines the (in)appropriate use of image tags in the whole process, according to the involved crowd. For example, in a scenario involving image tagging within a dating app, it could be the case that workers of diverse genders or regions view the appropriateness of tags relating to "attractiveness" differently.

Furthermore, KeepA(n)I is a valuable tool for researchers and educators examining the behavior of social media applications that rely on image tagging functionalities. It enables the auditing of tags applied to specific image sets for stereotypical patterns. Additionally, KeepA(n)I offers easy access to popular image taggers, facilitating the reverse engineering of applications that use these taggers for decision-making.

Limitations & Future Work

The KeepA(n)I platform is built as a proof-of-concept aiding stakeholders in auditing *AI-enabled Systems*, via descriptive and normative questions on potential stereotypes resulting from the incorporation of an image tagger in the System in question. As such, we recognize that this tool has the potential to be used with malice, aiding someone in promoting biases in the developed *System*. In the exploitation phase of the KeepA(n)I project, appropriate Terms of Service will be developed to prohibit such malicious uses of the platform.

An essential part of the platform is the evaluation of social norms that the *System* should respect, achieved by aggregating the perceptions of crowdworkers. We acknowledge that crowdworkers might suffer from cognitive biases (Draws et al. 2021) and through the design of the crowdsourcing task, as well as our cleaning and aggregation process, we aimed at minimizing this effect. Additionally, crowdsourced data can suffer from temporal variation (Christoforou, Barlas, and Otterbacher 2021), meaning that the perceptions on what tags are (in)appropriate might change over time and thus, the evaluation of social norms is not a "one-off". In this respect, the cost of evaluating the *System* for social norms is linked to many factors. As a future direction, we envision the inclusion of a consultation service within the KeepA(n)I platform, aimed at helping users strike a balance between cost-efficiency and accuracy in the auditing process.

Acknowledgments

This project has received funding from the Cyprus Research and Innovation Foundation under grant BRIDGE2HORIZON/0823E/0203 (PINNACLE) and EXCELLENCE/0421/0360 (KeepA(n)I), the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 739578 (RISE), the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

References

Baldassarre, M. T.; Caivano, D.; Fernandez Nieto, B.; Gigante, D.; and Ragone, A. 2023. The social impact of generative ai: An analysis on chatgpt. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, 363–373.

Barlas, P.; Krahn, M.; Kleanthous, S.; Kyriakou, K.; and Otterbacher, J. 2022. Shifting Our Awareness, Taking Back Tags: Temporal Changes in Computer Vision Services' Social Behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 22–31.

Barlas, P.; Kyriakou, K.; Guest, O.; Kleanthous, S.; and Otterbacher, J. 2021. To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3): 1–31.

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2019. Social b (eye) as: Human and machine descriptions of people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 583–591.

Christoforou, E.; Barlas, P.; and Otterbacher, J. 2021. It's about time: a view of crowdsourced data before and during the pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Clarifai. 2025. Clarifai, The World's Leading AI Lifecycle Platform. https://clarifai.com/. Accessed: 2025-04-14.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Ferrara, E. 2024. The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, 15: 100525.

Imagga. 2025. Image Recognition API, Computer Vision AI. https://imagga.com/. Accessed: 2025-04-14.

Judd, C. M.; and Park, B. 1993. Definition and assessment of accuracy in social stereotypes. *Psychological review*, 100(1): 109.

Kandori, M. 1992. Social norms and community enforcement. *The Review of Economic Studies*, 59(1): 63–80.

Marinucci, L.; Mazzuca, C.; and Gangemi, A. 2023. Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. *AI & SOCIETY*, 38(2): 747–761.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356.

Otterbacher, J. 2023. Computer Vision, Human Likeness, and Problematic Behaviors: Distinguishing Stereotypes from Social Norms. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 206–214.

Papakyriakopoulos, O.; Hegelich, S.; Serrano, J. C. M.; and Marco, F. 2020. Bias in word embeddings. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 446–457.

Steed, R.; and Caliskan, A. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 701–713.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Zlatev, J.; and Blomberg, J. 2019. Norms of language. *Normativity in language and linguistics*, 209: 69.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, in fact the KeepA(n)I approach promotes fairness by supporting non-expert developers and researchers in auditing AI systems.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
   (e) Did you describe the limitations of your work? No, because this is a demo paper and due to limited space it had to be omitted.
   (f) Did you discuss any potential negative societal impacts of your work? No, because this is a demo paper and due to limited space it had to be omitted.
   (g) Did you discuss any potential misuse of your work? No, because this is a demo paper and due to limited space it had to be omitted.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? No, due to limited space in this demo paper. We have provided a URL to the platform. In the platform under the My Campaigns feature, selecting Actions and then preview survey the user can view the crowdsourcing campaign shown to participants including the set of instructions given.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, all crowdsourcing experiments designed for the KeepA(n)I platform have received approval from the Cyprus National Bioethics Commitee. An information sheet together with an informed concent is presented to the participants and can be viewed in the platform (under the My Campaigns feature, selecting Actions and then preview survey).

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No, since according to the specific requests of a user, a crowdsouring campaign will be created dynamically and can have different durations. However, every campaign run in the KeepA(n)I platform rewards participants fairly providing at least the average hourly wage per country of residence of the participant and according to the crowdsourcing platform's suggestions.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, we briefly discuss how collected data, from a set of crowdsourcing campaigns, are presented to the user of the platform in section " Social Norms for an AI-enabled System". Additionally, we provide an image of how data are represented (see right image, Figure 1.)