9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Human Intuition as a Defense Against AI

Keywords: privacy protection, artificial intelligence, machine learning, link prediction, survey

The growing sophistication and ubiquity of AI-based invigilation tools is a persistent threat to the privacy of the general public. The increasing number of privacy-related scandals, such as Cambridge Analytica using the data of millions of Facebook users for political agendas, demonstrates just how vulnerable our private information is in the age of social media. Publicly available data, particularly social media data, can be used to uncover sensitive information. Most of the literature puts the responsibility of protecting the users' privacy in the hands of a centralized authority, even though such authority might be prone to error and negligence.

In this work, we evaluate the effectiveness of humans and AI in inferring and protecting private information. To this end, we focus on three attributes: (i) the *gender* of the author of a particular piece of text, (ii) the *location* in which a particular set of photos was taken, and (iii) the hidden *link* in a particular social network. For each of these attributes, we focus on two different tasks, and refer to the entity solving these tasks as an *agent*, which could be either a human or an AI. More specifically, in the first task, the attribute of interest is hidden, and the agent is required to infer this attribute from the given data. In the case of gender, for example, the agent is presented with a piece of text, and is required to infer the author's gender. In the second task, the attribute of interest is revealed, and the agent is required to modify the given data in order to make it harder for an algorithm to infer that attribute. In the case of gender, for example, the agent is presented with a piece of text along with the author's gender, and is asked to modify the text in order to keep author's gender hidden from prediction algorithms. The first task will be referred to as the *eye task*, since involves "seeing" hidden information, while the second task will be referred to as the *shield task*, since it involves "protecting" hidden information. The general outline of all eye and shield tasks is illustrated in Figure 1.

For the gender attribute, we generate the corresponding task instances using a dataset of Yelp reviews [3]. Each eye instance consists of the text of a review. To construct a shield instance, we select four words of the review that are most indicative of the author's gender (according to the normalized pointwise mutual information) as well as four words that are least indicative. The agent is asked to select which three of these eight words will be substituted with their synonyms. For the location attribute, we use a dataset of Flickr photos [5]. Each eye instance consists of a set of 16 random photos taken in the same country. To construct a shield instance, we identify four photos whose removal yields the largest drop in the location prediction accuracy, as well as four photos whose removal yields the smallest drop. The agent is asked to select which three of these eight photos will be removed. For the link attribute, we generate networks using Barabási-Albert [1], Erdős-Rényi [2], and Watts-Strogatz [4] models. Each eye instance consists of a network from which we randomly remove one link. To construct a shield instance, we identify two links whose removal causes the greatest decrease in the effectiveness of link prediction, as well as two links whose removal causes the least decrease. Moreover, we identify two links whose addition causes the greatest decrease, as well as two whose addition causes the least decrease. The agent is asked to select which three of these eight modifications will be introduced. We recruited participants using Amazon Mechanical Turk. Altogether, 1168 participants solved the comprehension test and completed their tasks.

Figure 2 compares the performance of humans and AI algorithms in the eye and shield tasks. As can be seen, the AI outperforms humans in every task. When we focus on the eye tasks, the average performance of both types of agents is the greatest in gender prediction,

followed by location prediction, with the average performance in link prediction being the poorest. This is consistent with the number of possible answers to each eye instance, as the agent has to choose one of two genders, one of twenty three countries, or one of seventy five possible links respectively. It is also worth noting that an average performance of both human and AI is at least as good as the random baseline in all eye tasks. When we turn our attention to the shields tasks, an average performance of AI is inversely proportional to the performance in the corresponding eye task. In other words, if the problem is difficult even without any strategic obfuscation, it makes it easier to add an additional layer of confusion. However, the situation is much more dire for the human agents attempting to perform a shield task. Their average performance oscillates around or even below (as in the case of links) the random baseline.

Figure 3 gives us a deeper insight into the differences in performance between the human and the AI agents. For about half the instances of the gender- and location-based eye tasks, the answers of human and AI agents are either both correct, or both incorrect. The gap in performance results from the distribution of the other half. Here, the number of instances in which AI, but not humans, give a correct answer is three to four times greater than those in which humans, but not AI give a correct answer. As for the link attribute, we observe a much greater percentage of instance where both types of agents give incorrect answers. Interestingly, when it comes to the instances in which the outcomes differ, we again see about threefold difference in AI's favor. These findings suggest that while the advantage of AI is considerable, there still exists a noteworthy class of instances in which humans have the upper hand. We now turn our attention to the shield tasks. Notice that for each of the three attributes, every instance has four of the possible modifications selected as having the greatest impact on prediction quality (we refer to these as the *high impact* answers), and the other four selected as having the smallest impact (we refer to these as the *low impact* answers). The figure shows that when trying to hide private information, people consistently select a smaller percentage of high impact answers than AI. The difference is especially pronounced in the case of location, as the inference algorithm simply takes a linear combination of the scores of all photos. Thus, an AI algorithm trying to obscure the location information has a relatively easy task of maximizing the score decrease. However, even for the other two attributes, where the prediction algorithms are non-linear, the ability of AI agents to identify high impact answers is much greater than human agents. This finding suggests that the inability of people to discern what aspects of the given instance are meaningful to the AI may be responsible for their poor performance.

Altogether, our results present a rather dire perspective on the human situation when it comes to the comparison with AI. Not only are human agents consistently outperformed by algorithms throughout all settings, but their abilities are particularly lacking when it comes to privacy protection. These findings underscore the need for both systemic solutions that guard our sensitive information, as well as tools and techniques that would assist us in taking the responsibility for protecting our privacy in our own hands.

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[2] P. Erdős and A. Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[3] S. Reddy and K. Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, 2016.

[4] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 1998.

[5] J. Yang, A. Chakrabarti, and Y. Vorobeychik. Protecting geolocation privacy of photo collections. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 524–531, 2020.

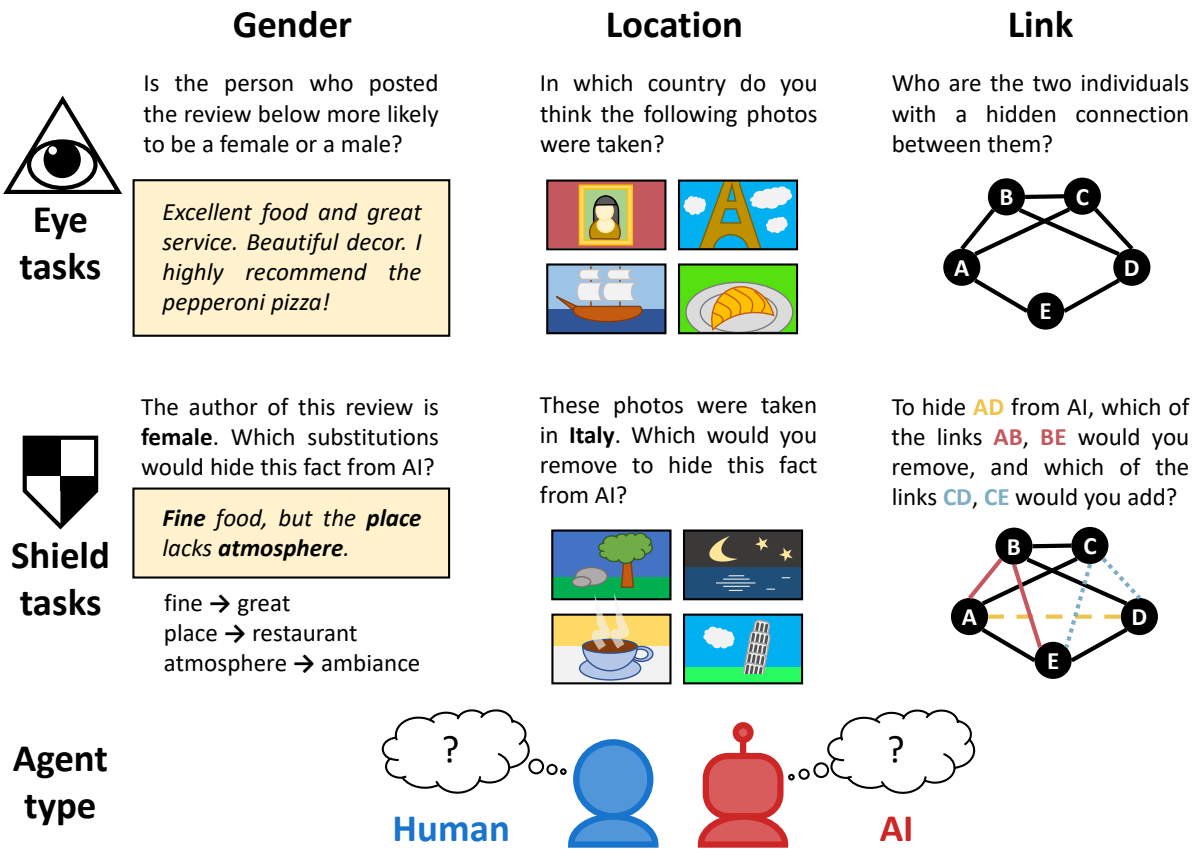

Figure 1: **The general outline of our experiment.** For each of the three attributes (gender, location, and link) we consider an *eye task* and a *shield task*. An eye task involves predicting the attribute based on available data: the gender of the author based on the text of the review, the country of origin based on a set of pictures, or the hidden connection based on the structure of a social network. A shield task involves modifying the data in order to make it more difficult for an AI to make the correct prediction. Each of the six tasks is given to two types of agents: to people (participants of an online survey) and to AI (algorithms trained on data) in order to compare their performance.

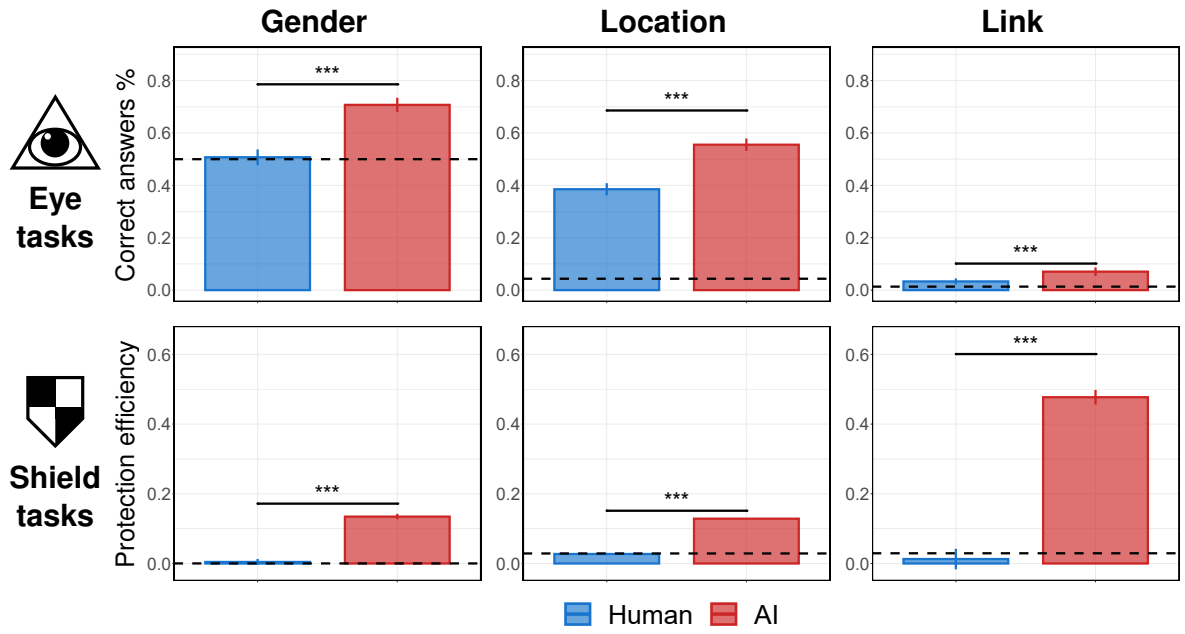

Figure 2: **Performance of people and algorithms in the eye and shield tasks.** Each column corresponds to a different attribute (gender, location, and link). The first row presents results of the eye tasks with y-axes corresponding to the percentage of the correct answers. The second row presents results of the shield tasks, with y-axes corresponding to the protection efficiency (negative values indicate that private information becomes more exposed). Each plot compares the average performance of survey participants and AI algorithms in a given task, with the dashed line marking the performance of a random baseline. Error bars represent 95% confidence intervals. All results are significant with p-values smaller than 0.001 according to the Welch's t-test.

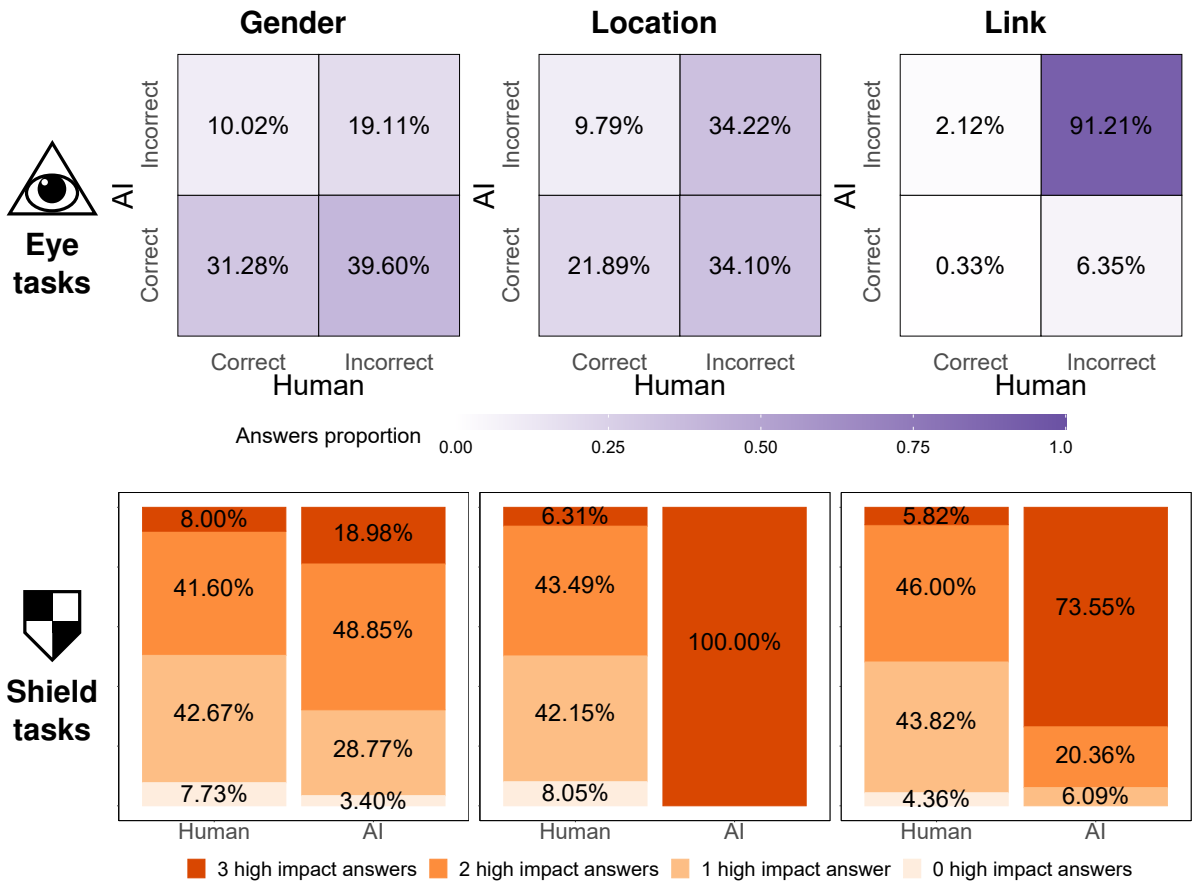

Figure 3: **Comparison of correct/incorrect and high/low impact answers in the eye and shield tasks.** The heatmaps in the first row compare the percentage of the eye instances correctly and incorrectly answered by the survey participants and by the AI. For any given instance, we consider the human answer to be correct if the majority of survey participants solved it correctly. For example, 31.28% in the bottom-left cell of the first heatmap indicates that, for 31.28% of the gender eye instances, both AI and over half of the survey participants gave a correct answer. The barplots in the second row compare the percentage of high and low impact answers selected by the survey participants and by the AI in the shields instances. For example, 42.67% in the first barplot indicates that, for the gender shield instances, 42.67% of the survey participants selected one high impact answer (and two low impact answers).