

Can AI-Assistance Improve Evaluations of Eyewitness Lineup Identifications?

Lauren E. Kelso
Charlottesville, VA
BA, University of Virginia, 2022

A Predissertation Research Project presented
to the Graduate Faculty of the
University of Virginia in
Candidacy for the Master of Arts

Department of Psychology
University of Virginia
August 2024

Readers:
Dr. Chad S. Dodson
Dr. Per B. Sederberg

Introduction

Eyewitness misidentifications are an important source of error in the legal system; they contribute to the majority of cases that have been later overturned by DNA evidence (Innocence Project, 2023). Yet, no tool exists that can assist law enforcement in distinguishing between reliable and unreliable eyewitnesses.

Once an eyewitness makes an identification from a lineup, standard police practice is to ask the witness to express how certain they are in their decision (Yates, 2017). Witnesses often prefer to express their confidence verbally (e.g. “I’m pretty sure”) as opposed to numerically (“80%”; e.g., Dodson & Dobolyi, 2015), but decades of research have shown that verbal expressions of confidence are often difficult for others to interpret (e.g., Beyth-Marom, 1982). Though previous work has shown that people can use an eyewitness’s testimony to judge the accuracy of their lineup identification (e.g., Smalarz & Wells, 2014), this ability is modest, and can be influenced by cognitive biases. Overall, it is important to identify a tool that can serve as a decision-aid and improve people’s evaluations of lineup identifications.

Artificial Intelligence (AI) assistance provides a promising method for producing this improvement. AI-assistance has been shown to improve people’s decision-making abilities on a variety of tasks (see Schemmer et al., 2022 for a meta-analysis), mainly because AI-predictions tend to be superior to human predictions (e.g., Alufaisan et al., 2021). Machine learning classifiers—trained solely on eyewitness verbal confidence statements—have been shown to accurately categorize out-of-sample lineup identifications as correct or incorrect approximately 75% of the time (Seale-Carlisle et al., 2022). So, providing people with a classifier’s prediction may lead to more accurate interpretations of eyewitness lineup identifications.

In this predissertation, I present two experiments that show that AI-assistance can improve people's evaluations of eyewitness lineup identifications. This paper includes a version of one first-authored paper (Kelso, Grabman, Dobolyi & Dodson, in press), and one soon to be submitted first-author manuscript. Part I shows that AI-assistance can eliminate a known cognitive bias—the Featural Justification bias. Whether or not this occurs, however, depends on the participant's perception of how useful they found the AI to be. Part II shows that in the absence of AI-assistance participants do show the ability to discriminate between correct and incorrect eyewitness lineup identifications. But, for feature-based and recognition-based lineup identifications, AI-assistance improves this ability. For familiarity-based identifications, however, discriminability was comparable for participants who received AI-assistance and those who did not. Altogether, these two studies suggest that AI-assistance can help people more accurately judge an eyewitness's lineup identification.

Introduction References:

- Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does Explainable Artificial Intelligence Improve Human Decision Making? In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, 6618-6626.
<https://doi.org/10.1609/aaai.v35i8.16819>
- Beyth-Marom, R. (1982). How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. *Journal of Forecasting*, 1(25), 257-269.
<https://doi.org/10.1002/for.3980010305>
- Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting Eyewitness Expressions of Confidence: The Featural Justification Effect. *Law and Human Behavior*, 39(3), 266-280. <https://doi.org/10.1037/lhb0000120>
- Innocence Project (2023). Eyewitness misidentification. *Innocence Project*.
<https://innocenceproject.org/eyewitness-misidentification/>.
- Kelso, L. E., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (in press). Does AI-assistance mitigate biased evaluations of eyewitness identifications? *Journal of Applied Research in Memory and Cognition*. <https://doi.org/10.1037/mac0000192>.
- Schemmer, M., Hemmer, P., Nitsche, M., Köhl, N., & Vössing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making.

- Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617-626.
<https://doi.org/10.1145/3514094.3534128>
- Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2022). The language of accurate and inaccurate eyewitnesses. *Journal of Experimental Psychology: General*, 151(6), 1283-1305. <https://doi.org/10.1037/xge0001152>
- Smalarz, L., & Wells, G. L. (2014). Post-Identification Feedback to Eyewitnesses Impairs Evaluators' Abilities to Discriminate Between Accuracy and Mistaken Testimony. *Law and Human Behavior*, 38(2), 194-202. <https://doi.org/10.1037/lhb0000067>.
- Yates, S.Q. (2017). Memorandum for Heads of Department Law Enforcement Components All Department Prosecutors. U.S. Department of Justice, Office of the Deputy Attorney General, 1-12. Retrieved from <https://www.justice.gov/file/923201/download>

Part I: Does AI-assistance mitigate biased evaluations of eyewitness identifications?

Journal of Applied Research in Memory and Cognition

Does Artificial Intelligence (AI) Assistance Mitigate Biased Evaluations of Eyewitness Identifications?

Lauren E. Kelso, Jesse H. Grabman, David G. Dobolyi, and Chad S. Dodson

Online First Publication, August 15, 2024. <https://dx.doi.org/10.1037/mac0000192>

CITATION

Kelso, L. E., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (2024). Does artificial intelligence (AI) assistance mitigate biased evaluations of eyewitness identifications?. *Journal of Applied Research in Memory and Cognition*. Advance online publication. <https://dx.doi.org/10.1037/mac0000192>

EMPIRICAL ARTICLE

Does Artificial Intelligence (AI) Assistance Mitigate Biased Evaluations of Eyewitness Identifications?

Lauren E. Kelso¹, Jesse H. Grabman², David G. Doboły³, and Chad S. Dodson¹

¹ Department of Psychology, University of Virginia, United States

² Department of Psychology, New Mexico State University, United States

³ Leeds School of Business, University of Colorado Boulder, United States



Artificial intelligence (AI) is playing an increasing role in human decision making. We use eyewitness lineup identification to show when AI assistance can and cannot help people avoid a cognitive bias known as the featural justification effect. People are biased to judge highly confident eyewitnesses as less likely to be correct when their lineup identification is based on an observable feature than on an expression of recognition. Our participants ($N = 1,010$) saw an eyewitness's lineup identification, accompanied by the eyewitness's verbal confidence statement (e.g., "I'm certain") and either a featural ("I remember his eyes") or a recognition justification ("I remember him"). They then rated the likely accuracy of the eyewitness's identification. AI assistance eliminated the featural justification effect but only in participants who regarded the AI as very useful. This project is the first step in evaluating human–algorithm interactions before the widespread use of AI assistance by law enforcement.


General Audience Summary

The use of artificial intelligence (AI) assistance is becoming increasingly common in many domains of human decision making, but there is no work investigating its use in the eyewitness domain. In this article, we used eyewitness lineup identification to show when AI assistance can and cannot help people avoid a cognitive bias known as the featural justification effect (FJE). The FJE distorts how people interpret an eyewitness's expression of confidence, leading them to discount the accuracy of a potentially correct perception of how useful they found the AI. For participants who found the AI to be highly useful, the FJE was eliminated. But for participants who did not find the AI to be useful, the FJE was robust. This project is a first step in evaluating how AI assistance might be useful for law enforcement before adoption by the legal system.

Keywords: explainable artificial intelligence, cognitive bias, eyewitness identification, cognitive forcing, artificial intelligence usefulness

Supplemental materials: <https://doi.org/10.1037/mac000192.supp>

Lauren E. Kelso  <https://orcid.org/0009-0002-2504-9313>

Jesse H. Grabman  <https://orcid.org/0000-0002-2526-1085>

David G. Doboły  <https://orcid.org/0000-0002-9493-3447>

The materials, data, and analysis scripts are available on Open Science Framework and can be accessed at <https://osf.io/ydgzh/> (Kelso et al., 2023). The study design, primary analysis plan, and predictions were preregistered on Open Science Framework (<https://osf.io/uxyzk/>) on November 17, 2023.

The authors have no conflicts of interest to declare. This research was supported by the National Science Foundation (Grant 2241989) awarded to

Chad S. Dodson. The funding source had no other involvement other than financial support.

No artificial intelligence-assisted technologies were used in this research or in the creation of this article. All participants were shown a consent form prior to starting the study. Only those who selected "Agree to continue" participated in the study. The University of Virginia institutional review board approved this research.

Lauren E. Kelso played a lead role in formal analysis, investigation, methodology, visualization, and writing—original draft. Jesse H. Grabman played a supporting role in methodology and writing—review and editing and an equal role in validation. David G. Doboły played a supporting role in

continued

From medical treatment to predicting recidivism, artificial intelligence (AI) is playing an increasingly larger role in human decision making (e.g., Day et al., 2018; McKinney et al., 2020; Pennisi et al., 2021). We use eyewitness lineup identification as a model paradigm to show when AI assistance can and cannot help people overcome a cognitive bias when judging the accuracy of an eyewitness's lineup identification.

After an eyewitness identifies someone from a lineup, law enforcement is advised to collect a confidence statement about the identification (Yates, 2017). Archival analyses of confidence statements show that roughly 30% of eyewitnesses justify their identification by referring to a visible feature of the suspect (Behrman & Richards, 2005). Likewise, at least 30% of mock witnesses refer to an observable feature when justifying their level of confidence in a lineup identification (Dobolyi & Dodson, 2018; Grabman et al., 2019). After collecting the confidence statement, police must assess the likely accuracy of the eyewitness's identification. However, there are a number of biases that can influence a person's interpretation of an eyewitness's lineup decision and confidence statement, such as the postidentification feedback effect (e.g., Smalarz & Wells, 2014).

One example of a cognitive bias that distorts how people interpret an eyewitness's lineup decision and confidence statement is the featural justification effect (FJE; Dodson & Dobolyi, 2015, 2017). When an eyewitness makes an identification from a lineup and refers to an observable feature in their confidence statement (e.g., "I am confident it's him. I remember his eyes"), they are perceived as less likely to be correct as compared with when an eyewitness's confidence statement is either recognition based (e.g., "I am confident it's him. I recognize him.") or consists of only an expression of confidence (e.g., "I am confident it's him") without an accompanying feature (e.g., Cash & Lane, 2017; Cash et al., 2024; Dobolyi & Dodson, 2018; Grabman et al., 2022). The FJE is strongest when the eyewitness makes their identification with a high level of confidence (e.g., 80%–100% confident; Dobolyi & Dodson, 2018). However, it is not the case that eyewitnesses are simply less accurate when they provide a featural rather than a recognition justification. In fact, all evidence indicates that eyewitness identifications are comparably likely to be correct when they include either a featural justification or a recognition justification (Dobolyi & Dodson, 2018; Grabman et al., 2019). In other words, the FJE reflects a misguided bias to discount the likely accuracy of a lineup identification when it is based on an observable feature. According to our perceived diagnosticity account, people discount the accuracy of such identifications because lineups typically consist of members who all look similar to each other. So, people are skeptical about the diagnostic value of a suspect's observable feature when this feature appears similar on all of the other faces in the lineup (e.g., Dodson & Dobolyi, 2015, 2017; see also Cash & Lane, 2017).

One method that may help people avoid the FJE is AI assistance. AI assistance improves people's decision making on many tasks (see Schemmer et al., 2022, for meta-analysis), in large part because


AI predictions are superior to human predictions (e.g., Alufaisan et al., 2021; Bansal et al., 2021; Buçinca et al., 2020; Fügener et al., 2021). Recent work (e.g., Grabman & Dodson, 2024; Seale-Carlisle et al., 2022) shows that machine learning classifiers can reliably distinguish between correct and incorrect eyewitness identifications on the basis of the eyewitness's verbal confidence statement and accompanying justification.


However, no one has investigated whether AI assistance can improve people's evaluation of an eyewitness's identification. Previous studies have shown that AI assistance can mitigate certain cognitive biases, but it can also exacerbate the effect of other biases (see Bertrand et al., 2022, for a review). A key variable when providing people with AI recommendations is how to convey the AI's advice to maximally improve decision making. An increasingly common method of presenting an algorithm's output is cognitive forcing functions. This method is designed to minimize a person's heuristic decision making and to maximize a more thoughtful consideration of the algorithm's advice (e.g., Buçinca et al., 2021; Eberhardt, 2020; Green & Chen, 2019). The cognitive forcing method typically uses a two-step procedure in which individuals first make a prediction based on raw information about the task but without AI assistance. They then are shown the AI's advice and are given the opportunity to revise their prediction. For example, Green and Chen (2019) showed across separate tasks that people made more accurate loan default and recidivism predictions in a cognitive forcing condition relative to a typical one-step condition where they received the AI's estimate together with the raw information (e.g., a narrative profile) and had to make a prediction. Another method of presenting AI information is explainable AI (XAI). In contrast to simply providing the user with the AI's output, XAI provides a person with the algorithm's advice, along with an explanation of how the AI evaluates the importance of individual features. Ideally, XAI allows the person to better understand the basis for the algorithm's output, which may improve their decision making (e.g., Buçinca et al., 2021). There are inconsistent findings about the type of assistance that is most beneficial. Some studies suggest that providing explanations can improve performance above and beyond simply presenting the AI's advice (e.g., Buçinca et al., 2020), but this benefit has not been observed in other studies (e.g., Alufaisan et al., 2021). In short, the best way to present AI assistance remains an open question.


The Present Study

Will AI assistance improve people's evaluation of eyewitness identifications by minimizing the FJE? We chose the FJE to test AI assistance for two reasons. First, the FJE has been shown to be a strong cognitive bias with respect to effect size (i.e., large effect size observed by both Cash & Lane, 2017, and Dodson & Dobolyi, 2017). Consequently, if AI assistance can mitigate a strong bias, then we can be optimistic about generalizing the effects of AI assistance to other cognitive biases. Second, the FJE illustrates an

methodology, software, and writing—review and editing and an equal role in validation. Chad S. Dodson played a lead role in conceptualization, funding acquisition, and writing—review and editing and a supporting role in validation.

 The data are available at <https://osf.io/ydgzh/>.

 The experimental materials are available at <https://osf.io/ydgzh/>.

 The preregistered design and analysis plan are accessible at <https://osf.io/uxyzk>.

Correspondence concerning this article should be addressed to Lauren E. Kelso, Department of Psychology, University of Virginia, P.O. Box 40040, Charlottesville, VA 22904-4400, United States. Email: lek9kx@virginia.edu

interdisciplinary problem of what causes people to misunderstand verbal probability statements (e.g., Beyth-Marom, 1982; Budescu et al., 2014). Recently, there is a growing interest among eyewitness researchers about whether evaluators accurately interpret eyewitness verbal statements of confidence (e.g., Greenspan & Loftus, 2024; Smalarz et al., 2021), and the FJE is an example of a biased evaluation of an eyewitness confidence statement.

All participants saw a series of trials, each involving an eyewitness's identification from a lineup that was accompanied by the eyewitness's confidence statement. This confidence statement included either a featural or a recognition justification. Participants also received either no AI assistance (control condition) or AI assistance, which took one of three forms. They saw either the AI's prediction about the likely accuracy of the identification (Prediction Only condition), the AI's prediction as well as a graphical explanation (Prediction + Graphical Explanation condition), or they were in a Cognitive Forcing condition. Participants then rated the likely accuracy of the eyewitness's identification.

We predict that we will replicate the FJE in the No AI assistance condition. Specifically, we will observe lower perceived accuracy ratings of identifications that are accompanied by featural statements than recognition statements. In the conditions that receive AI assistance, we expect that participant perceptions of AI usefulness will be a moderating variable. In other domains, an individual's perception of a tool's usefulness can be a strong predictor of how they use the tool (e.g., Egelman et al., 2008; Venkatesh et al., 2003). In our study, we expect that those who find the AI less useful will be less likely to consider its prediction when making their accuracy rating.

Our central prediction is that AI assistance will minimize and possibly eliminate the FJE, particularly in participants who rate the AI as more (vs. less) useful. Specifically, we predict that participants who find the AI's predictions to be more useful will be more likely to align their perceived accuracy ratings about the eyewitness's identification to the AI's prediction than will those who find the AI to be less useful. Because the AI's predictions are comparable for the featural and recognition statements if participants follow the AI's advice, they will rate both types of statements similarly, thus overcoming the FJE.

Our final prediction is that participant's accuracy ratings will align more closely with the AI's predictions when evaluating recognition statements than featural statements. Contrary to existing research about AI assistance, we expect more resistance to the AI's advice about featural statements—so less of an effect on the FJE—in the Cognitive Forcing and the Prediction + Graphical Explanation conditions than in the Prediction Only condition. We predict that the colorization and highlighting of particular words (e.g., haircut) in the former two conditions will draw participants' attention to these words, which should increase the likelihood of activating the featural bias.

Method

We preregistered our predictions, design, and analysis plan on Open Science Framework and can be accessed at <https://osf.io/uxyzk>.

Participants

Our final sample consisted of 1,010 participants (52.25% female, 76.63% White/Caucasian) between the ages of 18 and 94 ($M = 45.10$, $SD = 13.52$) who were recruited through Amazon Mechanical Turk

(MTurk) in exchange for compensation. We excluded participants who showed attempts at taking the survey multiple times (e.g., duplicate worker IDs), had a VPN that was from outside of the United States, failed our initial attention check, failed the colorblindness check (see the Procedure section), answered "yes" to whether they had seen any of the faces in the study before, or indicated they had technical errors that interfered with their ability to complete the task. This sample translated to over 120 participants in each of our eight between-subjects conditions. A priori power analyses deemed our sample size sufficient to detect a medium-sized effect at an α of .05 with over 99% power. The University of Virginia institutional review board approved this research.

Design

This study used a 2 (Statement Type: Recognition, Featural) \times 4 (AI Assistance: None, Prediction Only, Prediction + Graphical Explanation, Cognitive Forcing) between-subjects design. The dependent variable was perceived accuracy ratings.

Materials

All stimuli were from a lineup paradigm conducted by Grabman et al. (2019) and consisted of responses from mock witnesses who (a) identified someone from a lineup, (b) provided a typed verbal expression of confidence and self-reported either a featural or a recognition justification, and (c) then provided a high numeric level of confidence (80%–100%).

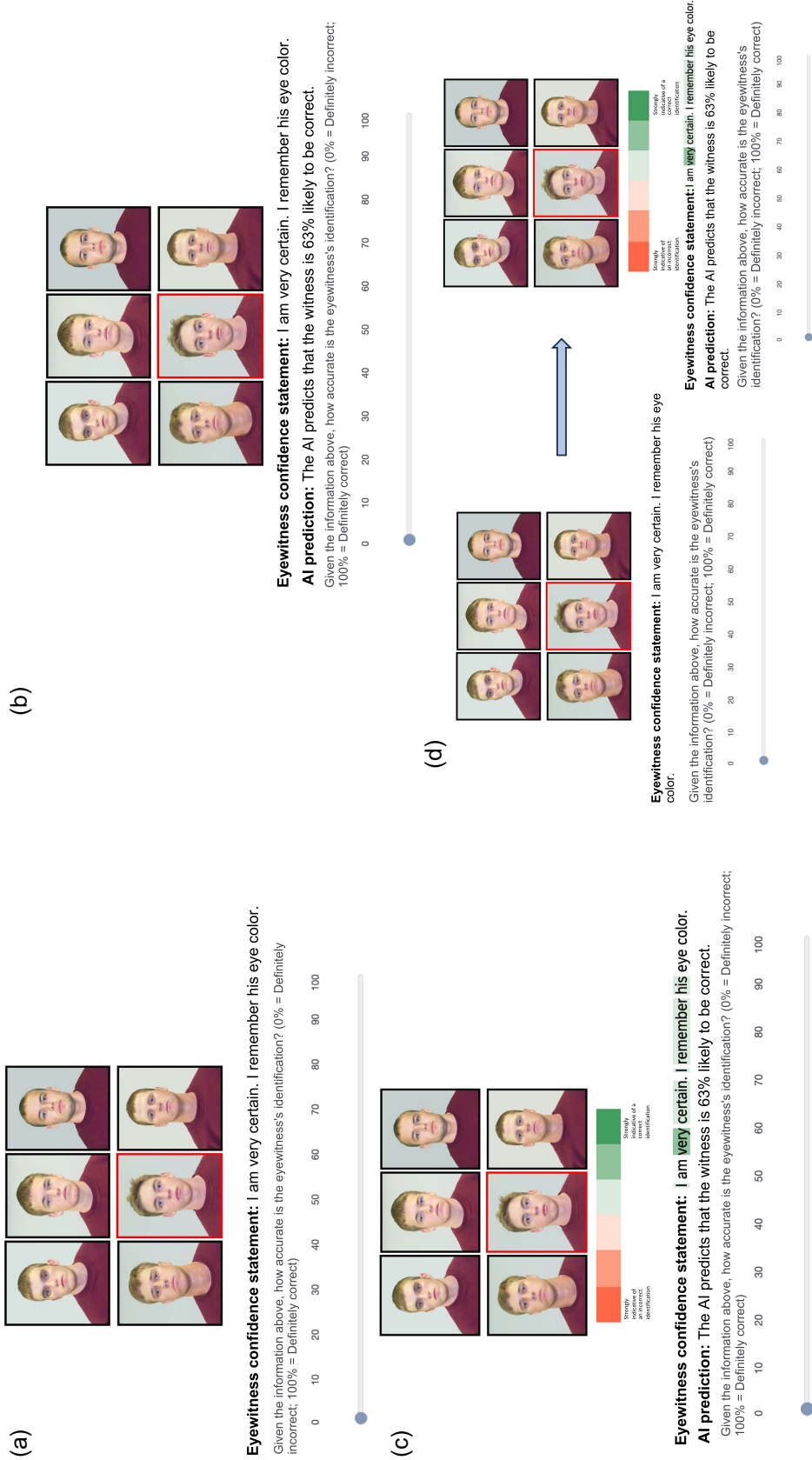
Lineups

Participants evaluated confidence statements provided to a total of six unique lineups. As shown in Figure 1, each lineup consisted of six White males, wearing matching, maroon-colored t-shirts, arranged in a 2 \times 3 array. In addition to the lineup, participants in our study saw who the mock witness in Grabman et al.'s (2019) paradigm had selected from the lineup. This person was outlined in red, while all other lineup members were outlined in black. Although our participants did not know which response was correct, all lineup responses were correct identifications (i.e., the "target" was chosen) because we needed a large pool of high confidence identifications.

Confidence Statements

We used only identifications that were made at high levels of numeric confidence (80%–100% confident). Average numeric confidence was comparable for featural statements ($M = 93.17\%$, $SD = 9.60\%$) and recognition statements ($M = 96.25\%$, $SD = 7.89\%$). Though participants did not see the witness's numeric confidence level, we chose high-confidence identifications because the FJE is strongest when the witness is highly confident (e.g., Dobolyi & Dodson, 2018). Each lineup decision was accompanied by a statement of confidence that was generated by the participant–witness in Grabman et al.'s (2019) paradigm. We selected two different kinds of statements: recognition statements (e.g., "I am certain. I remember him.") and featural statements (e.g., "I am certain. I remember his eyes."). We generated a pool of 89 statements (48 recognition, 41 featural) from which we randomly

Figure 1
The Four Artificial Intelligence Assistance Conditions



Note. Panel A shows the No AI assistance condition, Panel B shows the Prediction Only condition, Panel C shows the Prediction + Graphical Explanation condition, and Panel D shows the Cognitive Forcing condition. AI = artificial intelligence. See the online article for the color version of this figure.

sampled six statements that were shown to each participant (see the Supplemental Material for the individual confidence statements).

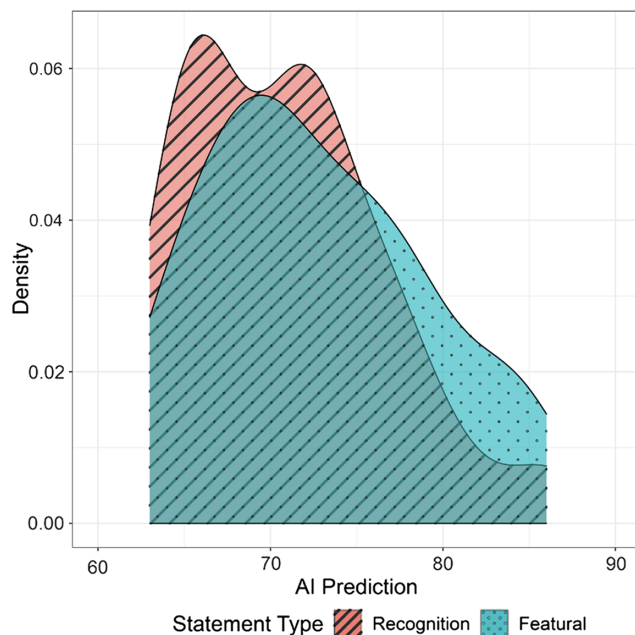
Artificial Intelligence Prediction

Participants in our three AI assistance conditions saw an AI's prediction of the likely accuracy of the eyewitness's identification. These predictions come from a least absolute shrinkage and selection operator logistic classifier developed by Seale-Carlisle et al. (2022) on Grabman et al.'s (2019) statements. The classifier had been trained to use the words in the entire corpus of confidence statements from Grabman et al. (2019) and showed a strong ability to distinguish between correct and incorrect identifications. Its prediction about the accuracy of a particular identification is based on the kind and number of words with diagnostic value in the corresponding confidence statements. Although all statements in this study were associated with a correct identification, there was a range of classifier probabilities for the featural and recognition statements. Figure 2 shows that the range of AI-predicted probabilities for the recognition and featural statements extends from 63% to 86%. In addition, the overall average AI prediction was comparable for the recognition statements ($M = 70.95\%$, $SD = 5.77\%$) and the featural statements ($M = 72.72\%$, $SD = 6.41\%$; see the Supplemental Material for the classifier prediction for each statement).

Artificial Intelligence Usefulness

Participants in all conditions, aside from the No AI assistance condition, answered three questions about the usefulness of the AI's prediction in helping them judge the accuracy of the eyewitness's

Figure 2
Density Plot Showing the Distribution of Artificial Intelligence Predictions Across Both Statement Types



Note. AI = artificial intelligence. See the online article for the color version of this figure.

Figure 3

The Three Artificial Intelligence Usefulness Questions Answered by Participants Who Interacted With the Artificial Intelligence

In evaluating the performance of the AI-generated predictions in correctly judging the accuracy of the eyewitnesses, I believe that the tool was:

Not helpful at all	○ ○ ○ ○ ○ ○	Very helpful
Not valuable at all	○ ○ ○ ○ ○ ○	Very valuable
Not useful at all	○ ○ ○ ○ ○ ○	Very useful

Note. For example, those in the Prediction Only, Prediction + Graphical Explanation, or Cognitive Forcing conditions. AI = artificial intelligence. See the online article for the color version of this figure.

identification (Abbasi et al., 2021; Venkatesh et al., 2003). Figure 3 shows the three AI Usefulness questions that participants answered. They answered these questions with a 6-point Likert scale, which ranged from 1 (*not helpful/valuable/useful at all*) to 6 (*very helpful/valuable/useful*). We computed an "AI Usefulness" score for each participant, which was the average of their responses to the three questions. Figure 4 shows the distribution of AI Usefulness scores in the present study ($M = 3.72$, $SD = 1.48$).

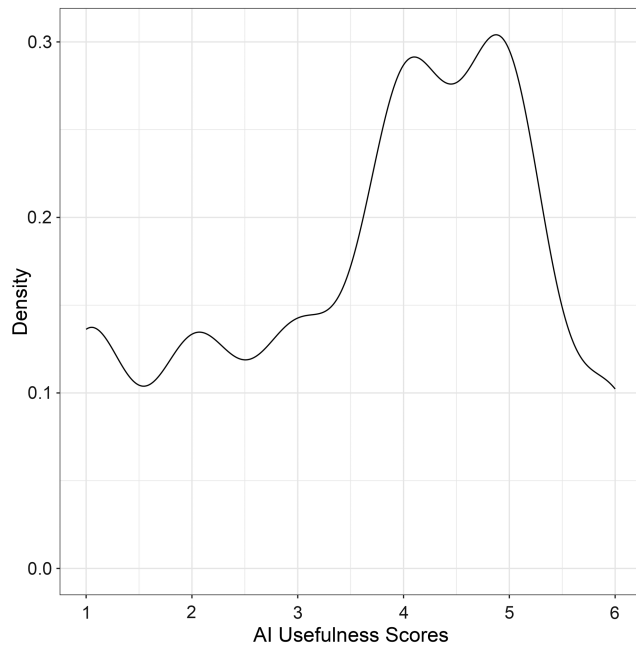
Procedure

Because two of our AI assistance conditions used red/green colorization, we administered an Ishihara color blindness test (Wellcome Library & London, 2018) to all participants. Only those who passed the color blindness check continued on with the study. Participants were then instructed to imagine that they were police officers, and their job was to judge the likely accuracy of eyewitness identifications. They were told that they would see the exact same lineup as the eyewitness and that the person the witness had selected from the lineup would be highlighted in red. Participants were also told that each lineup would be accompanied by the eyewitness's written expression of certainty in their lineup decision. Finally, they were informed that the witnesses they were rating were chosen at random and that "it is possible that [they] will see all inaccurate witnesses, all accurate witnesses, or some combination of accurate and inaccurate witnesses." To demonstrate the task, participants saw a lineup of six colorful smiley faces, one highlighted in red, and an accompanying confidence statement which read, "I know it's him. I remember that his face was green." Participants were instructed to rate the witness as "definitely correct" (i.e., slide the bar to 100%).

Figure 1 shows an example of the task in each of the four AI assistance conditions. In all four conditions, participants saw a six-person lineup, with the person selected by the mock witness from Grabman et al.'s (2019) paradigm outlined in red. They also saw the mock witness's typed statement of confidence in their identifications. Participants either saw all featural or all recognition confidence statements. In all conditions, participants rated the likely accuracy of the eyewitness's identification using a 101-point scale, which ranged from 0% (*definitely incorrect*) to 100% (*definitely correct*).

As depicted in Panel A, participants in the No AI assistance (control) condition relied solely on their own impressions of the witness's accuracy. Participants in the three conditions that received AI assistance—Panels B, C, and D—received instructions about seeing an AI's prediction of the likely accuracy of the eyewitness's

Figure 4
Density Plot Showing the Distribution of Artificial Intelligence Usefulness Scores



Note. AI = artificial intelligence.

identification. As a brief background about how the AI generates the predictions, participants were told that:

The AI considers each individual word in the eyewitness's statement, evaluating if the word is more indicative of a correct identification or an incorrect identification. The AI then uses that information to make a prediction about the likely accuracy of the witness.

Panel B of Figure 1 shows the Prediction Only condition. In this condition, participants additionally received the AI's prediction about the likely accuracy of the mock witness (e.g., "The AI predicts this witness is 63% likely to be correct").

Participants in the Prediction + Graphical Explanation (Panel C) and the Cognitive Forcing (Panel D) conditions were shown additional instructions, providing them additional insights into the underlying source of the AI's prediction. They were told that "some of the words in the witness's statement of confidence will be highlighted in varying shades of green and red" and were shown a color legend. As shown in Panel C, green shading was more indicative of a correct identification, and red shading was more indicative of an incorrect identification. Additionally, the darkness of the color indicated the importance of the word in the AI's prediction.

Finally, Panel D of Figure 1 shows the Cognitive Forcing condition. This condition consisted of two parts. Part I was equivalent to the No AI assistance condition (i.e., participants rated the likely accuracy of the witness after seeing only the lineup, lineup decision, and statement). In Part II, participants were shown the equivalent of the Prediction + Graphical Explanation condition and given the opportunity to update their accuracy rating.

In each condition, participants completed six trials. After completing all trials, participants in the Prediction Only, Prediction +

Graphical Explanation, and Cognitive Forcing conditions answered the AI Usefulness questions. Finally, all participants were asked to complete a short demographic survey that included questions about age, sex, and race.

Results

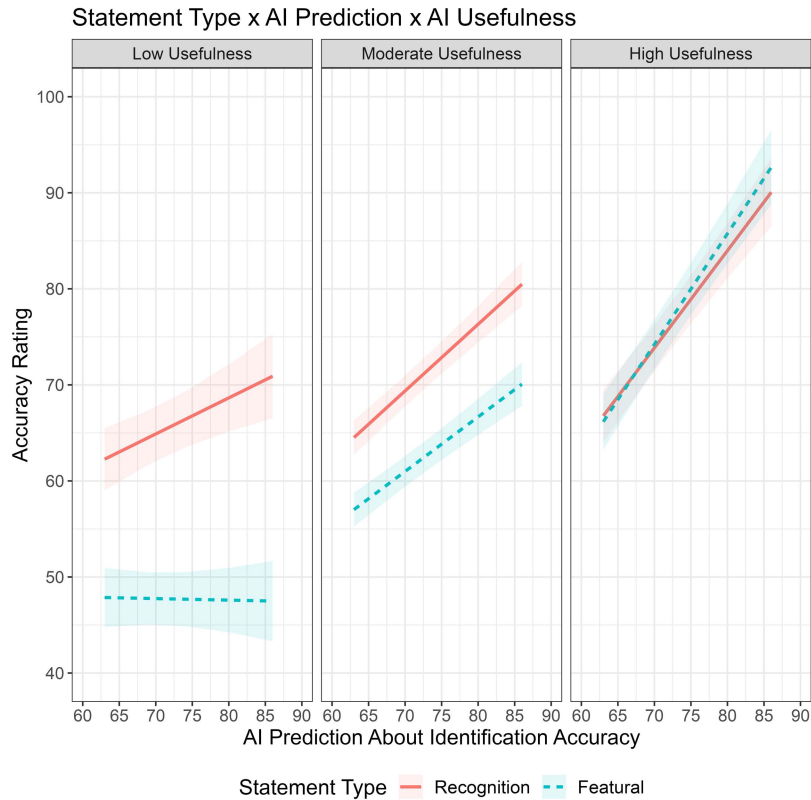
Does AI assistance minimize the FJE? Before answering this question, we examined performance in the No AI assistance (control) condition. Consistent with our prediction of observing an FJE, accuracy ratings were lower for featural statements ($M = 55.52$, $SD = 16.93$) than for recognition statements ($M = 65.83$, $SD = 20.68$), $t(246.77) = 4.37$, $p < .001$.

To answer the question of whether AI assistance reduces the FJE, we used a linear mixed-effects model to predict accuracy ratings from the fixed factors of AI assistance (Prediction Only, Prediction + Graphical Explanation, Cognitive Forcing), Statement Type (Recognition, Featural), AI Prediction, and AI Usefulness with random intercepts of participant and lineup. In Wilkinson and Rogers' (1973) notation, the model was Accuracy Ratings \sim Statement Type \times AI assistance \times AI Prediction \times AI Usefulness + (1|Participant) + (1|Lineup). Our model excluded the No AI assistance condition because participants in this condition did not interact with the AI; therefore, we were unable to collect AI Usefulness scores from them. For the Cognitive Forcing condition, our model only included participant's accuracy ratings from Part II (i.e., after they had the opportunity to update their rating). We used the *lme4* package (Bates et al., 2015) in R v.4.3.2. (R Core Team, 2023) to fit our model to the data. We used the *car* package (Fox & Weisberg, 2019) to obtain the analysis of variance table for our model. Finally, we assessed the absolute fit for our model. Using the MuMIn package (Bartón, 2023), we calculated both the conditional (R2GLMM(c)) and marginal (R2GLMM(m)) pseudo-R2 for fixed effects. The conditional pseudo-R2 considers the variance accounted for by the random effects in the model, and the marginal pseudo-R2 only considers the variance from fixed effects. Our model adequately fits the data, pseudo-R2GLMM(m) = 0.21; pseudo-R2GLMM(c) = 0.54. Overall, this model is based on 4,530 responses from 755 participants.

The model showed main effects of AI assistance $\chi^2(2) = 16.62$, $p < .001$; Statement Type $\chi^2(1) = 75.35$, $p < .001$; AI Prediction $\chi^2(1) = 368.31$, $p < .001$; and AI Usefulness $\chi^2(1) = 142.43$, $p < .001$. There were also significant two-way interactions between Statement Type and AI Usefulness $\chi^2(1) = 36.03$, $p < .001$; AI assistance and AI Prediction $\chi^2(2) = 10.71$, $p = .005$; and AI Usefulness and AI Prediction $\chi^2(1) = 60.76$, $p < .001$. Additionally, there was a significant three-way interaction between Statement Type, AI Usefulness, and AI Prediction $\chi^2(1) = 5.51$, $p = .019$. We begin by discussing the three-way interaction, which moderates all lower-level effects, apart from the AI Assistance \times AI Prediction interaction.

Figure 5 shows the interaction between Statement Type, AI Prediction, and AI Usefulness. Although AI Usefulness was a continuous factor in the analysis, for ease of interpretation, the three panels in Figure 5 show participant perceptions of AI usefulness split into tertiles. The "Low Usefulness" tertile includes AI Usefulness scores of 2.4 or lower, "Moderate Usefulness" includes scores higher than 2.4 but less than or equal to 3.9, and "High Usefulness" includes scores higher than 3.9. In all panels the red (solid) and blue (dashed) lines refer to responses to the recognition statements and featural statements, respectively. A difference in

Figure 5
Three-Way Interaction Between Statement Type, Artificial Intelligence's Prediction, and Artificial Intelligence Usefulness



Note. Colored lines represent model estimates and error shading indicates the 95% confidence interval. Panels show participant's AI usefulness scores split into tertiles (Low: ≤ 2.4 , Moderate: > 2.4 and ≤ 3.9 , High: > 3.9). AI = artificial intelligence. See the online article for the color version of this figure.

perceived accuracy ratings for the recognition and featural statements reflects the FJE.

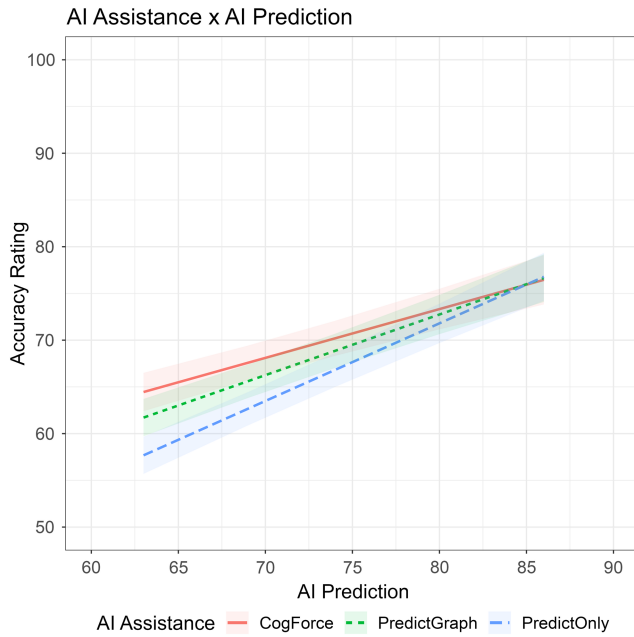
Figure 5 shows two main results that are the basis for the interaction. First, and most importantly, the magnitude of the FJE depends on participant's perception of the AI's usefulness. When participants do not find the AI useful (i.e., the leftmost panel), the FJE is robust; accuracy ratings are substantially lower for featural statements than for recognition statements. However, the magnitude of the FJE steadily diminishes with increasing perceptions of AI usefulness, as illustrated by the decreasing gap in perceived accuracy ratings between featural and recognition statements across the three panels. The FJE completely disappears when participants find the AI to be very useful (i.e., the rightmost panel). As we predicted, participants who find the AI tool to be very useful rate the recognition and featural statements comparably, effectively eliminating the featural justification bias.

The second key result in Figure 5 is that when participants find the AI less (vs. more) useful, they are more resistant to using the AI's prediction when they rate the likely accuracy of featural than recognition statements. This result is visible by the greater change in the slope for the featural than for the recognition statements across the three panels. For example, participants who do not find the AI

useful (i.e., leftmost panel) are completely resistant to the AI's prediction when judging the accuracy of featural statements, as illustrated by the flat line for the featural statements. With increasing perceptions of AI usefulness—see moderate and high usefulness panels—participants are increasingly likely to use the AI's prediction as a basis for their accuracy ratings for both the featural statements and the recognition statements.

Figure 6 shows the two-way interaction between the type of AI assistance the participant received and the AI's prediction. The influence of a particular type of AI assistance depends on the magnitude of the AI's prediction, such that when the AI's prediction is low, participant accuracy ratings are highest in the Cognitive Forcing condition and lowest in the Prediction Only condition. However, as the AI's prediction increases, participants in all three AI assistance conditions increase their accuracy ratings, with eventual convergence of the three group's ratings at the highest end of the AI's predictions. Additionally, Figure 6 suggests that participants are most responsive to AI assistance in the Prediction Only condition—and least responsive in the Cognitive Forcing condition—as shown by the difference in slopes between the three AI assistance groups. The slope for the Prediction Only condition is steeper than for either the Cognitive Forcing or Prediction + Graphical Explanation conditions,

Figure 6
Two-Way Interaction Between Artificial Intelligence Assistance Condition and Artificial Intelligence's Prediction



Note. Colored lines represent model estimates, and error shading indicates the 95% confidence interval. AI = artificial intelligence; CogFor = Cognitive Forcing condition; PredictGraph = Prediction + Graphical Explanation condition; PredictOnly = Prediction Only condition. See the online article for the color version of this figure.

indicating that participants in the Prediction Only condition are adjusting their accuracy ratings more as the AI's prediction increases than those in the other two conditions.

Discussion

Problems with eyewitness identification are an important source of error in the legal system (Garrett, 2017). No tool exists that can minimize the influence of various contextual effects that can adversely affect people's understanding of an eyewitness's verbal expression of confidence and, consequently, their predictions about eyewitness identification accuracy. A classifier that serves as a decision aid for law enforcement may help to reduce some of the problems with eyewitness identification.

Can AI assistance help people overcome a cognitive bias? When judging the accuracy of a highly confident eyewitness's lineup identification, people are biased against and perceive an eyewitness as less likely to be correct when their lineup identification is based on a visible feature (e.g., "I remember his eyes") than when it is based on a recognition response (e.g., "I remember him")—a bias that we call the FJE. Consistent with previous studies (e.g., Cash & Lane, 2017; Dobolyi & Dodson, 2018; Dodson & Dobolyi, 2015; Grabman et al., 2022), participants in our control (No AI assistance) condition showed the featural justification bias and rated identifications as less likely to be correct when they were accompanied by featural statements than recognition statements.

Our key novel finding is that we show that AI assistance can eliminate the featural justification bias. But whether or not this occurs depends on participants' perception of AI usefulness; this bias is eliminated in participants who rate the AI as very useful, but it is robust in participants who distrust the AI. These results highlight the necessity of collecting participant perceptions of AI usefulness when evaluating the influence of AI assistance on people's behavior. Previous work has shown that how useful participants find a tool to be influences the way that tool is used (e.g., Egelman et al., 2008; Venkatesh et al., 2003), and our findings further support this result.

We also predicted that participants would be more resistant to considering the AI's advice when evaluating featural than recognition statements. In addition, we predicted that this resistance would be more pronounced in the Cognitive Forcing and the Prediction + Graphical Explanation conditions because we thought that the colorization of particular words in these conditions would exacerbate the featural justification bias. These predictions were only partially supported. Though we did find more resistance to the AI's advice when judging featural statements than recognition statements, this resistance was present in all conditions involving AI assistance. Contrary to our prediction, highlighting particular words was not more likely to activate resistance to judging featural statements. Furthermore, Figure 6 suggests that participants were least responsive to the AI's advice in the Cognitive Forcing condition and most responsive in the Prediction Only condition, as shown by the difference in the slope of the lines for each AI assistance condition. In other words, as the AI's prediction increases, participants show more of an adjustment of their perceived accuracy ratings when only presented with the AI's prediction as compared with when they are presented with the AI's prediction and a graphical explanation. Altogether, our results indicate that all three types of AI assistance are comparably effective at mitigating the FJE.

Because there is no other (to our knowledge) research on the topic of AI assistance and eyewitness lineup identifications, our study leaves many questions open for future research. One important question is whether AI assistance can improve people's ability to discriminate between correct and incorrect eyewitness identifications. Additionally, due to material constraints, our AI's predictions were restricted to a range of 63%–86%. There is the potential that participants respond quite differently to an AI's prediction when it is higher or lower than the range in our study. Overall, we are not arguing for the immediate adoption of AI assistance by law enforcement. Before that can happen, we need greater confidence that our AI predictions about identification accuracy—which are based on laboratory paradigms—scale up and generalize to real-world eyewitness lineup identifications.

In sum, for the first time, we show that AI assistance can help people overcome a cognitive bias known as the FJE. Receiving AI assistance, relative to receiving no assistance, helps people to evaluate the accuracy of an eyewitness's identification better but only when they find the AI to be useful. If they do not find the AI to be useful, they are unlikely to consider its advice when making their decision. The rapid spread of classifier (AI) assistance across various domains of human decision making means that there is a need to investigate its use in the eyewitness domain before adoption by the legal system.

References

- Abbasi, A., Dobolyi, D., Vance, A., & Zahedi, F. M. (2021). The phishing funnel model: A design artifact to predict user susceptibility to phishing websites. *Information Systems Research*, 32(2), 410–436. <https://doi.org/10.1287/isre.2020.0973>
- Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision making? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6618–6626. <https://doi.org/10.1609/aaai.v35i8.16819>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). *Does the whole exceed its parts? The effect of AI explanations on complementary team performance* [Conference session]. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3411764.3445717>
- Bartón, K. (2023). *MuMIn: Multi-model inference* (R package Version 1.47.5). <https://CRAN.R-project.org/package=MuMIn>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, 29(3), 279–301. <https://doi.org/10.1007/s10979-005-3617-y>
- Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). *How cognitive biases affect XAI-assisted decision-making: A systematic review* [Conference session]. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom. <https://doi.org/10.1145/3514094.3534164>
- Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting*, 1(3), 257–269. <https://doi.org/10.1002/for.3980010305>
- Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). *Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems* [Conference session]. Proceedings of the 25th International Conference on Intelligent User Interfaces, New York, NY, United States. <https://doi.org/10.1145/3377325.3377498>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188. <https://doi.org/10.1145/3449287>
- Budescu, D. V., Por, H. H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, 4(6), 508–512. <https://doi.org/10.1038/nclimate2194>
- Cash, D. K., & Lane, S. M. (2017). Context influences interpretation of eyewitness confidence statements. *Law and Human Behavior*, 41(2), 180–190. <https://doi.org/10.1037/lhb0000216>
- Cash, D. K., Russell, T. D., Harrison, A. T., & Papesh, M. H. (2024). Evaluating eyewitnesses: Translating expressions of pre-and post-identification confidence. *Applied Cognitive Psychology*, 38(1), Article e4163. <https://doi.org/10.1002/acp.4163>
- Day, M. Y., Cheng, T. K., & Li, J. G. (2018). *AI robo-advisor with big data analytics for financial services* [Conference session]. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain. <https://ieeexplore.ieee.org/document/8508854>
- Dobolyi, D. G., & Dodson, C. S. (2018). Actual vs. perceived eyewitness accuracy and confidence and the featural justification effect. *Journal of Experimental Psychology: Applied*, 24(4), 543–563. <https://doi.org/10.1037/xap0000182>
- Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior*, 39(3), 266–280. <https://doi.org/10.1037/lhb0000120>
- Dodson, C. S., & Dobolyi, D. G. (2017). Judging guilt and accuracy: Highly confident eyewitnesses are discounted when they provide featural justifications. *Psychology, Crime & Law*, 23(5), 487–508. <https://doi.org/10.1080/1068316X.2017.1284220>
- Eberhardt, J. L. (2020). *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. Penguin Books.
- Egelman, S., Cranor, L. F., & Hong, J. (2008). *You've been warned: An empirical study of the effectiveness of web browser phishing warnings* [Conference session]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy. <https://doi.org/10.1145/1357054.1357219>
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). SAGE Publications. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become Borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly*, 45(3), 1527–1556. <https://doi.org/10.25300/MISQ/2021/16553>
- Garrett, B. L. (2017). Convicting the innocent redux. In D. S. Medwed (Ed.), *Wrongful convictions and the DNA revolution: Twenty-five years of freeing the innocent* (pp. 40–56). Cambridge University Press. <https://doi.org/10.1017/9781316417119.004>
- Grabman, J. H., Cash, D. K., Slane, C. R., & Dodson, C. S. (2022). Improving the interpretation of verbal eyewitness confidence statements by distinguishing perceptions of certainty from those of accuracy. *Journal of Experimental Psychology: Applied*, 28(3), 589–605. <https://doi.org/10.1037/xap0000362>
- Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*, 8(2), 233–243. <https://doi.org/10.1037/h0101835>
- Grabman, J. H., & Dodson, C. S. (2024). Unskilled, underperforming, or unaware? Testing three accounts of individual differences in metacognitive monitoring. *Cognition*, 242, Article 105659. <https://doi.org/10.1016/j.cognition.2023.105659>
- Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 50. <https://doi.org/10.1145/3359152>
- Greenspan, R. L., & Loftus, E. L. (2024). Interpreting eyewitness confidence. Numeric, verbal, and graded verbal scales. *Applied Cognitive Psychology*, 38(1), Article e4151. <https://doi.org/10.1002/acp.4151>
- Kelso, L. E., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (2023). *AI-assistance and the featural justification effect* (Open Science Framework). <https://osf.io/ydgzh/>
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- Pennisi, M., Kavasidis, I., Spampinato, C., Schinina, V., Palazzo, S., Salaniti, F. P., Bellitto, G., Rundo, F., Aldinucci, M., Cristofaro, M., Campioni, P., Pianura, E., Di Stefano, F., Petrone, A., Albarello, F., Ippolito, G., Cuzzocrea, S., & Conoci, S. (2021). An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. *Artificial Intelligence in Medicine*, 118, Article 102114. <https://doi.org/10.1016/j.artmed.2021.102114>
- R Core Team. (2023). *A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). *A meta-analysis of the utility of explainable artificial intelligence in*

- human-AI decision-making* [Conference session]. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. <https://doi.org/10.1145/3514094.3534128>
- Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2022). The language of accurate and inaccurate eyewitnesses. *Journal of Experimental Psychology: General*, *151*(6), 1283–1305. <https://doi.org/10.1037/xge0001152>
- Smalarz, L., & Wells, G. L. (2014). Post-identification feedback to eyewitnesses impairs evaluators' abilities to discriminate between accurate and mistaken testimony. *Law and Human Behavior*, *38*(2), 194–202. <https://doi.org/10.1037/lhb0000067>
- Smalarz, L., Yang, Y., & Wells, G. L. (2021). Eyewitnesses' free-report verbal confidence statements are diagnostic of accuracy. *Law and Human Behavior*, *45*(2), 138–151. <https://doi.org/10.1037/lhb0000444>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *Management Information Systems Quarterly*, *27*(3), 425–478. <https://doi.org/10.2307/30036540>
- Wellcome Library, London. (2018). *L0059158 Eight Ishihara charts for testing colour blindness, Europe* [Photograph]. https://commons.wikimedia.org/wiki/File:Eight_Ishihara_charts_for_testing_colour_blindness,_Europe_Wellcome_L0059158.jpg
- Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, *22*(3), 392–399. <https://doi.org/10.2307/2346786>
- Yates, S. Q. (2017). *Memorandum for heads of department law enforcement components all department prosecutors*. U.S. Department of Justice, Office of the Deputy Attorney General. <https://www.justice.gov/file/923201/download>

Received March 7, 2024

Revision received June 27, 2024

Accepted June 28, 2024 ■

Part II: When can AI-assistance improve people's ability to distinguish between correct and incorrect eyewitness lineup identifications?

Introduction

Mistaken eyewitness identification is one of the leading causes of false convictions (National Registry of Exonerations, 2023). Yet, no tool exists that assists law enforcement with distinguishing between reliable and unreliable eyewitness identifications. We show, for the first time, that the assistance of artificial intelligence (AI) can improve people's decision-making and increase their ability to discriminate correct from incorrect lineup identifications.

After an eyewitness identifies someone from a lineup, standard police practice is to ask them to express their level of certainty in the identification in their own words (Yates, 2017). This practice is common in the United States and many countries have similar guidelines (Fitzgerald, Rubinova, & Juncu, 2021). In general, eyewitnesses prefer to express confidence in words (e.g., "I'm pretty sure") as opposed to numbers (e.g., "80%"; Behrman & Richards, 2005; Dodson & Dobolyi, 2015), but regardless of whether confidence is registered verbally or numerically, higher (vs. lower) confidence is associated with greater identification accuracy (Wixted & Wells, 2017; Smalarz et al., 2021).

In addition to expressing their level of certainty, many eyewitnesses include a justification for their identification (Behrman & Richards, 2005). Archival analyses show that approximately 30-50% of real-world eyewitnesses justify their lineup decision by referring to at least one visible feature of the suspect (e.g., "I remember his eyes"; Behrman & Richards, 2005; Steblay & Wells, 2023). Similar frequencies have been observed in laboratory studies that use mock-witness paradigms (Dobolyi & Dodson, 2018; Grabman et al., 2019). However, laboratory studies also show that eyewitnesses frequently make recognition-based (e.g., "I remember him") and familiarity-based (e.g., "He looks familiar") identifications—approximately 27% and 23% of the time, respectively, in Grabman et al. (2019; see also Dobolyi & Dodson, 2018). Altogether,

extant research shows that references to facial features and expressions of recognition/recollection and familiarity are the most common bases for an eyewitness's lineup identification.

Why does the basis of the eyewitness's lineup identification matter? One reason is that the eyewitness's confidence-accuracy relationship is much weaker for familiarity-based than for feature- and recognition-based identifications (e.g., Dobolyi & Dodson, 2018; Grabman et al., 2019). Whereas high (vs. low) confidence identifications are strongly predictive of accuracy for feature- and recognition-based identifications, this is not the case for familiarity-based identifications. Moreover, reality monitoring research shows that true memories are often accompanied by more frequent recollection of sensory and perceptual details as compared to false memories (e.g., Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981; Schooler et al., 1986). So, eyewitness statements that refer to specific facial features or recollective experiences should be more likely to be correct than statements that simply refer to a feeling of familiarity. Consequently, evaluators should be better able to predict the accuracy of eyewitness identifications that are feature- and recognition-based than those that are familiarity-based.

How well can people predict the accuracy of an eyewitness's identification? Smalarz and Wells (2014) showed participants videotaped testimony from four eyewitnesses who had made either a correct or an incorrect lineup identification. Participants then made a binary decision of whether they believed the witness was accurate or not. Their participants could distinguish accurate from inaccurate witnesses; they believed accurate witnesses approximately 70% of the time, while only believing inaccurate witnesses approximately 36% of the time (see also Beaudry et al., 2015; Grabman, Dobbins, & Dodson, 2024; Kaminski & Sporer, 2017 for similar findings). Though people do show some ability to discriminate between accurate and inaccurate

witnesses, discriminability is often far from perfect (e.g., average accuracy of 67.2% in Smalarz & Wells, 2014). Furthermore, modest discriminability has also been observed in basic memory (i.e., non-eyewitness) paradigms in which people are asked to classify another's memory as true or false (i.e., accuracy rate of 56.0% in Gamoran et al., 2024; see also Clark-Foos, Brewer, & Marsh, 2015; Schooler et al. 1986). Overall, people exhibit some ability to determine when another's memory is correct or incorrect, but it is clear that there is room for improvement.

One tool that may improve discriminability is the assistance of AI. From predicting recidivism to deciding whether someone will default on their loan, AI-assistance has improved people's decision-making on a variety of tasks (see Schemmer et al., 2022 for a meta-analysis). This improvement is largely attributed to the fact that AI predictions tend to be superior to human predictions (e.g., Alufaisan et al., 2021; Bansal et al., 2021; Buçinca et al., 2020; Fügener et al., 2021). For example, Seale-Carlisle and colleagues (2022) demonstrated that machine learning classifiers, trained only on eyewitness verbal confidence statements about a lineup identification, can reliably categorize out-of-sample lineup identifications as either correct or incorrect approximately 75% of the time (see also Grabman & Dodson, 2024; Grabman et al., 2024; Dobbins & Kantner, 2019 for similar findings)—a level of discrimination accuracy that is higher than what has been observed by human participants. Because classifiers are performing better than humans in many tasks, presenting people with classifier predictions may improve their ability to distinguish between correct and incorrect eyewitness lineup identifications.

However, one critical question about AI-assistance is what the best way is to convey the AI's prediction to human participants. One frequently-used method involves presenting participants with the task information (e.g., a loan applicant profile) and then showing them the AI's prediction (e.g., "This person will default on their loan"; Wang et al., 2022). Another

method, known as Explainable AI (XAI), involves providing participants the task information, showing the AI's prediction, and giving participants an explanation of how the AI came to its conclusion (e.g., showing how the AI weighted the different factors in a loan applicant's profile). Ideally, XAI allows participants to understand the basis for the AI's prediction, making them more likely to detect errors (e.g., Buçinca et al., 2021) and further improves their discriminability beyond what could be achieved from presenting participants with the AI's advice alone. While some studies have found support for XAI as being more beneficial at improving people's discriminability as compared to just providing the AI advice alone (e.g., Buçinca et al., 2020), others have found little difference between the two ways of presenting AI output (e.g., Alufaisan et al., 2021; Kelso et al., in press).

Finally, an individual difference factor that could influence how a person uses the AI's advice is their perception of the usefulness of the AI's prediction. Previous studies have shown that how useful people find a tool to be is a strong predictor of how they will use the tool (e.g., Abbasi et al., 2021; Venkatesh et al., 2003; Egelman et al., 2008; Kelso et al., in press), as those who do not find it useful are less likely to consider its recommendations.

Current Study

Will AI-assistance improve people's ability to discriminate between correct and incorrect eyewitness identifications? To answer this question, we showed participants a series of eyewitness lineup identifications, accompanied by the eyewitness's statement of confidence. The confidence statement included both a verbal expression of certainty, such as "I'm pretty sure," and either a featural, recognition, or familiarity justification. Participants either received no assistance from the AI (Control condition) or AI-assistance, which took one of two forms: (1) Prediction Only—participants saw the AI's prediction about whether the eyewitness was correct

or incorrect, or (2) Prediction + Graphical Explanation—participants saw a graphical explanation along with the AI’s prediction. All participants judged whether the eyewitness’s identification was correct or incorrect.

We pre-registered two predictions. First, we predicted that receiving AI-assistance would improve participants’ ability to discriminate between correct and incorrect identifications, as compared to the no assistance control condition. Second, we expected that participants’ perceptions of AI usefulness would moderate the effectiveness of AI-assistance. Specifically, and analogous to what we observed in Kelso et al. (in press), we predicted that participants who perceive the AI as more (vs. less) useful would show better discriminability.

Moreover, although not pre-registered, we expected that people would show better discriminability when they evaluate eyewitness identifications that are either feature- or recognition-based than when they are familiarity-based. This expectation is based on the body of work (e.g., Grabman et al., 2019) showing that eyewitness confidence is a stronger predictor of accuracy for feature- and recognition-based identifications than it is for familiarity-based identifications.

Method

We preregistered our design and predictions on Open Science Framework (OSF):

<https://osf.io/mcfz9>

Participants

We recruited participants through Amazon Mechanical Turk (mTurk), who completed the study in exchange for monetary compensation. Participants were excluded from analyses if they (1) showed attempts at taking the survey multiple times (e.g., duplicate Worker Ids), (2) were from outside of the United States, (3) failed either the initial attention check or the color

blindness check (see Procedures below), (4) indicated that they had seen our stimuli prior to this experiment, or (5) stated that they had technical errors that interfered with their ability to complete the task. Our final sample consisted of 1092 participants (54.98% Female, 77.31% White/Caucasian) between the ages of 19 and 87 ($M = 43.62$, $SD = 12.66$). This provided us with approximately 120 participants in each of our nine between-subjects conditions. A-priori power analyses deemed our sample size sufficient to detect a medium sized effect at an alpha of 0.05 with over 99% power. The University of Virginia Institutional Review Board (IRB) approved this research.

Design

This study's design consisted of two between-subjects factors – Statement Type (Recognition, Featural, Familiarity) and AI-assistance (None, Prediction Only, Prediction + Graphical Explanation) – and one within-subjects factor of Trial Accuracy (Correct, Incorrect).

Materials

The experiment used stimuli from a lineup paradigm conducted by Grabman et al. (2019). These stimuli included responses from mock witnesses who (a) chose someone from a lineup, (b) provided a typed verbal expression of confidence for their decision, and (c) provided a numeric level of confidence (0-100%). As explained in this previous paper's methods,

independent coders slotted justifications into several categories, including those of interest in the present work (featural, recognition, or familiarity).

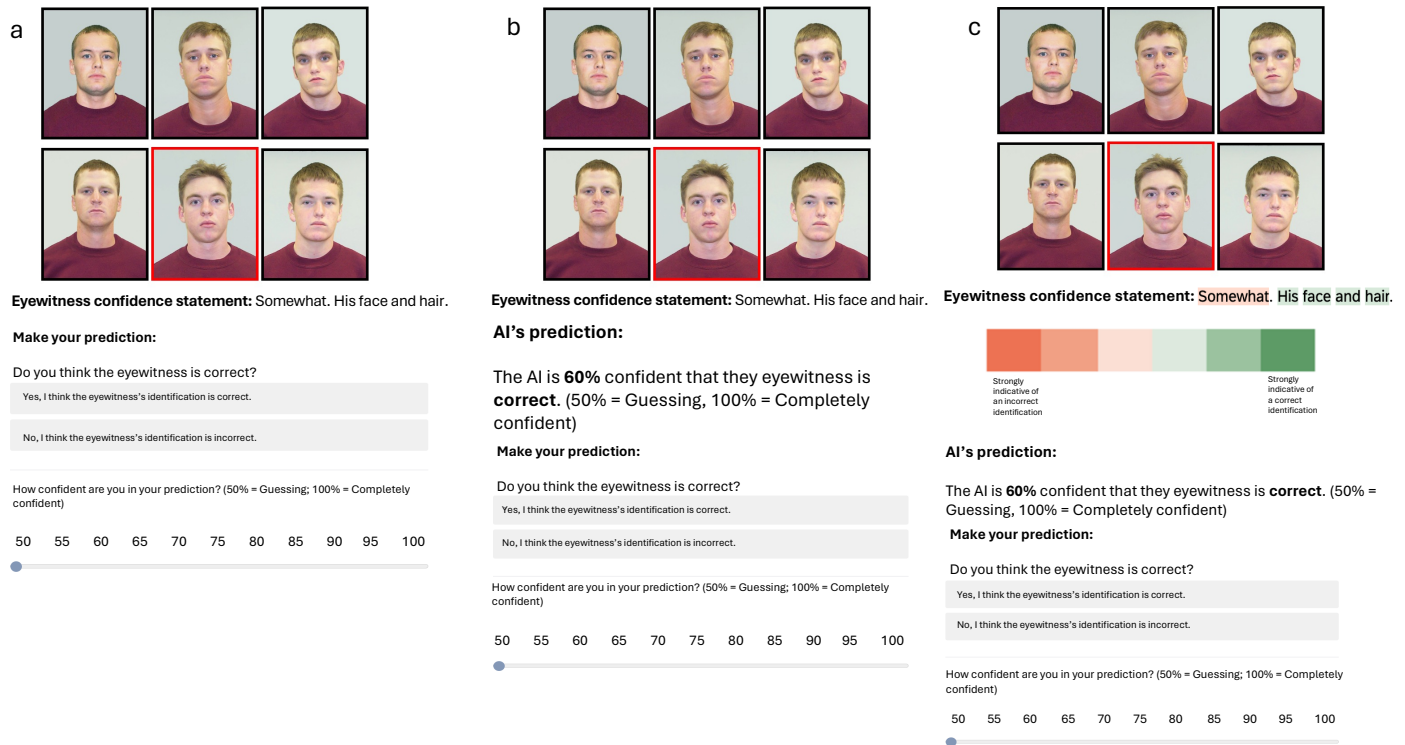


Figure 1. The three AI-assistance conditions. Panel A shows the No-assistance condition, panel B shows the Prediction Only condition, and Panel C shows the Prediction + Graphical Explanation condition.

Lineups

All participants assessed mock-witnesses' confidence statements that were associated with one of six unique lineups. As shown in Figure 1, each lineup was arranged in a 2 x 3 array and showed six white males. The person who the mock-witness in Grabman et al. (2019) selected was outlined in red and all other lineup members were outlined in black. Participants

saw a total of six different lineups and the mock-witness's identification was correct in half of them and in the remaining lineups the identification was incorrect.

Confidence Statements

Each lineup decision was accompanied by a confidence statement that was provided by the mock-witnesses in Grabman et al. (2019). Each confidence statement consisted of a verbal expression of certainty and a justification. We selected three kinds of statements: recognition (e.g., "I am certain. I remember him."), featural (e.g., "I am certain. I remember his eyes.") and familiarity (e.g., "I am certain. He looks familiar"). This provided us a pool of 875 confidence statements (326 recognition, 303 featural, 246 familiarity). For each participant, we randomly sampled six statements from each of their assigned statement type's pool, with the constraint that the same lineup could not be used more than once (See the supplemental material for the individual confidence statements). Though participants were not presented with information about the eyewitness's numeric level of confidence, mock-witnesses also provided such a rating (0-100%, in increments of 20%) for each identification. Table 1 shows the frequency of numeric confidence ratings for each statement type. The mean numeric confidence rating was comparable for feature-based ($M = 68.95\%$, $SD = 24.98\%$) and recognition-based ($M = 64.83\%$, $SD = 29.81\%$) identifications. But, average numeric confidence was substantially lower for familiarity-based ($M = 43.74\%$, $SD = 25.21\%$) identifications.

EW Confidence Level	Statement Type		
	<i>Featural</i>	<i>Recognition</i>	<i>Familiarity</i>
<i>0</i>	13	68	99
<i>20</i>	166	315	712
<i>40</i>	360	312	517
<i>60</i>	440	386	451
<i>80</i>	711	543	314
<i>100</i>	500	566	79

Table 1. Frequency of lineup identifications at each level of numeric confidence.

AI Predictions

In addition to the lineup, lineup identification, and mock-witness's statement, participants in the Prediction Only and Prediction + Graphical Explanation conditions were also shown an AI's prediction of whether the eyewitness's identification was correct or incorrect and the level of confidence in its prediction. The predictions and confidence judgements come from a least absolute shrinkage and selection operator (LASSO) logistic regression classifier that was developed by Seale-Carlisle et al. (2022) and trained on the entire corpus of confidence statements from Grabman et al. (2019). The classifier provides a probability score, ranging from 0%-100%, that the identification is correct. Probabilities less than 50% correspond to a prediction that the identification is incorrect whereas probabilities greater than 50% correspond to a prediction that the identification is correct. Probabilities closer to the endpoints (i.e., 0% or 100%) reflect increasing confidence in the prediction and probabilities closer to 50% reflect uncertainty. This classifier distinguished between correct and incorrect identifications, with an Area Under the Curve (AUC) of .77 (chance performance is an AUC of 0.50).

For each confidence statement, we first identified whether the classifier predicted that the identification was correct or incorrect (i.e., any probability above 50% was labeled as "correct").

Next, we specified the classifier’s level of confidence on a scale from 50% (guessing) to 100% (completely confident). For ‘correct’ predictions, the confidence score was simply the associated probability (i.e., a probability of 60% was presented as the “The AI is 60% confident that the eyewitness’s identification is correct”). For ‘incorrect’ predictions, in order to translate the classifier’s confidence onto the same 50% - 100% scale, we subtracted the probability level from 100% (i.e., a classifier probability of 40%—which signifies an incorrect response—was converted to 60% [i.e., $100\% - 40\% = 60\%$], e.g., “The AI is 60% confident that the eyewitness is incorrect”).

AI Usefulness

In both AI-assistance conditions, participants answered three questions that gauged their perceptions of how useful the AI was in assisting their decision-making (Kelso et al., in press). Figure 2 shows the three questions that participants answered. All three questions were answered on a 6-point Likert scale which ranged from 1 (Not helpful/valuable/useful at all) to 6 (Very helpful/valuable/useful). For each participant, we computed an “AI Usefulness” score, which was the average of their responses to these three questions ($M = 3.83$, $SD = 1.42$).

In evaluating the performance of the AI-generated predictions in correctly judging the accuracy of the eyewitnesses, I believe that the tool was:

Not helpful at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very helpful
Not valuable at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very valuable
Not useful at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very useful

Figure 2. The three AI usefulness questions answered by participants in the Prediction Only and Prediction + Graphical Explanation conditions.

Procedure

Because our Prediction + Graphical Explanation condition used red/green colorization, we administered an Ishihara color blindness test (Wellcome Library, London, 2018) to all participants. Only participants who passed the color blindness check were allowed to begin the experiment. Our initial set of instructions told participants to imagine that they were a police officer, and that their job was to decide whether an eyewitness's lineup identification was accurate or inaccurate. They were informed that they would see the identical lineup as the eyewitness, that the person that the witness had selected from the lineup would be outlined in red, and that the witness's statement of confidence would accompany the lineup. Finally, to avoid strategic guessing (Smalarz & Wells, 2014), participants were told that, "It is possible that [they] will see all inaccurate, all accurate, or some combination of accurate and inaccurate witnesses." As a demonstration of the task, and as a final attention check, all participants saw a lineup which consisted of six colorful smiley faces. The green smiley face was highlighted in red and was accompanied by a confidence statement which read, "I know it's him. I remember that his face was green." Participants were instructed to select "Yes, I think the eyewitness is correct" and to indicate their confidence in this selection. Only participants who selected that the witness was correct continued with the study.

Figure 1 shows an example of the task in each of the three AI-assistance conditions. In all conditions, participants saw a lineup, the eyewitness's lineup selection, and the eyewitness's accompanying statement of confidence. Participants saw either all featural, all recognition, or all familiarity statements. Participants were asked to provide their prediction of the witness's accuracy by answering a two-alternative forced choice question where they decided whether they believed the witness's identification was correct or incorrect. After making their decision,

participants provided their confidence in their decision on a 51-point scale, which ranged from 50% (Guessing) to 100% (Completely confident).

In the No assistance (Control) condition, shown in Panel A, participants saw the lineup, the witness's selection (outlined in red), and the witness's confidence statement. Participants based their prediction and confidence solely on their own impressions of the witness's accuracy.

Participants who received AI-assistance (panels B and C) were provided with additional instructions. This set of instructions provided a brief background about the AI and how it generates its predictions. Participants were told that, "The AI considers each individual word in the eyewitness's statement, evaluating if the word is more indicative of a correct identification or an incorrect identification. The AI then uses that information to make a prediction about the likely accuracy of the witness."

In the Prediction Only condition, shown in panel B, participants received the AI's binary prediction about whether the witness was correct or incorrect (e.g., "The AI predicts that the witness is correct") and the AI's confidence in its decision (e.g., "The AI is 60% confident in its prediction). In the Prediction + Graphical Explanation (Panel C) condition, participants were shown an additional set of instructions. These instructions informed participants that, "some of the words in the witness's statement of confidence will be highlighted in varying shades of green and red" and were shown a color legend. The legend was comprised of three different shades of red and three of green. Darker green shading was more indicative of a correct identification and darker red shading was more indicative of an incorrect identification. Aside from the colorization of the text, the Prediction + Graphical Explanation condition was identical to the Prediction Only condition.

All participants completed a total of six trials; in three trials they were shown correct eyewitness identifications and in three trials they saw incorrect identifications. Participants in the Prediction Only and the Prediction + Graphical Explanation conditions answered the AI Usefulness questions after completing all six trials. Finally, all participants completed a short demographic survey.

Results

Does receiving AI-assistance improve peoples' ability to discriminate between correct and incorrect eyewitness identifications? To measure discrimination, we first translated the combination of each participant's predictions and confidence ratings onto a continuous scale from 0-100 (0 = Definitely incorrect, 100 = Definitely correct). For example, if a participant answered "Yes, I think the eyewitness is correct" with 80% confidence, their score would be 80. Conversely, if a participant answered "No, I think the eyewitness is incorrect" with 80% confidence, their score would be 20. We converted our participants' responses to a continuous scale because this enables the construction of Receiver-Operating Characteristic (ROC) curves. Though this is a deviation from our preregistered analysis plan of using a linear mixed-effects model, constructing ROC curves is a more appropriate analysis as ROC provides a non-parametric method of examining discriminability which controls for participants' overall bias toward believing (or not believing) the witness. We constructed ROC curves using the *pROC* package (Robin et al., 2011) in R v.4.3.3 (R Core Team, 2024). This package also provided an

Area Under the Curve (AUC) value for each curve, with higher AUC values indicating better discrimination ability (chance performance is an AUC of .50).

We predicted that AI-assistance would help people discriminate between correct and incorrect eyewitness identifications. To analyze performance between conditions, we used the DeLong method (DeLong, DeLong & Clarke-Pearson, 1988) to compare the respective group's AUC scores. Because we are performing multiple comparisons, we used the *stats* package (R Core Team, 2024) to adjust the *p*-values using the Benjamini-Hochberg method to control for the false discovery rate (Benjamini & Hochberg, 1995).

To illustrate performance in the different conditions, Figures 3-5 show the rate at which participants responded “Yes, I believe the eyewitness is correct” to trials in which the eyewitness's identification either was correct (the blue bars) or incorrect (the orange bars; Smalarz & Wells, 2014). To contextualize this figure, perfect discrimination is shown by a 100% rate of responding “Yes” to correct trials (i.e., blue bars) and a 0% rate of responding “Yes” to incorrect trials (i.e., orange bars). In contrast, equivalent rates of responding “Yes” to correct and incorrect trials would indicate a complete lack of discrimination. We calculated the mean and 95% confidence interval for each group using the *plotrix* (Lemon, 2006) package. Each AI-assistance condition's corresponding mean AUC value is shown at the top of Figures 3-5; for comparison, we also show the performance of our AI model alone (i.e., the far-right column of bars within Figures 3-5 [labeled “AI”]).

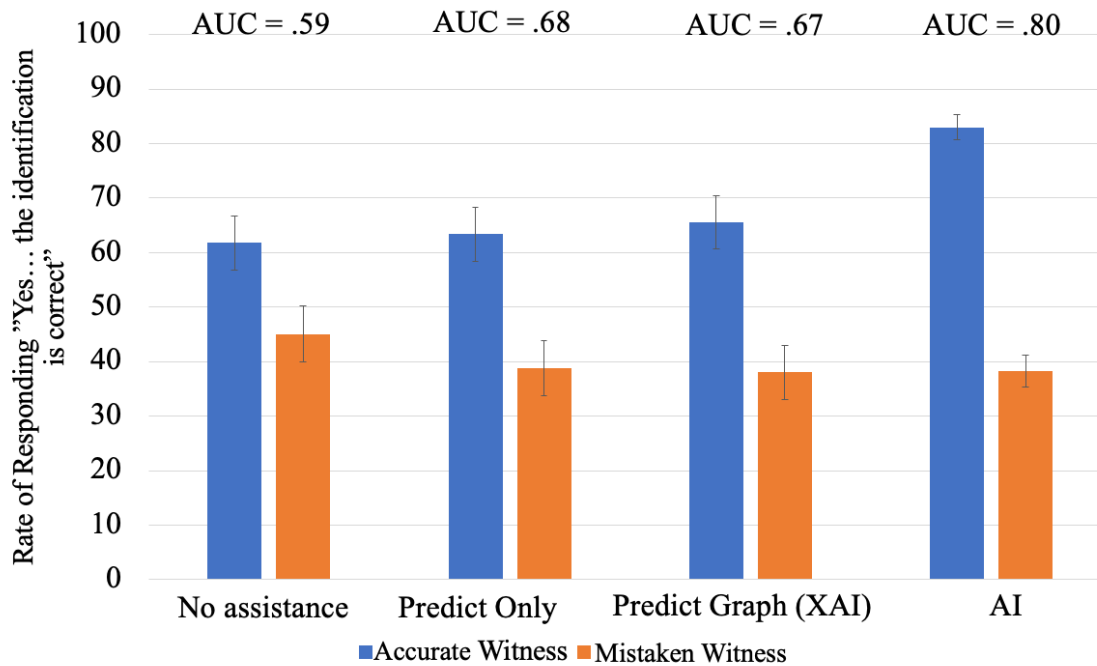


Figure 3. For feature-based identifications, the rate of responding “Yes, I think the eyewitness’s identification is correct” to accurate and mistaken eyewitness identification trials in the different AI-assistance conditions. Errors bars indicate a 95% confidence interval.

Feature-Based Identifications. Figure 3 shows the rate at which participants believed feature-based identifications were correct (i.e., rate of responding “Yes I think the eyewitness’s identification is correct”) in the different AI-assistance conditions. In the No-assistance (control) condition participants discriminated between correct and incorrect feature-based identifications at above chance levels (AUC = .59; 95% CI [.55-.64]). Both AI-assistance conditions significantly improved discrimination as compared to the No-assistance condition: No-assistance vs. Prediction Only (AUC = .68; 95% CI [.64-.72]), $D = 3.06$, $p = .006$; No-assistance vs. Prediction + Graphical Explanation (AUC = .67; 95% CI [.63-.71]), $D = 2.49$, $p = .025$. Discrimination was similar between the Prediction Only and Prediction + Graphical Explanation conditions, $D = 0.55$, $p = .698$. The far-right bars in Figure 3 shows discrimination performance of the AI alone (AUC = .80; 95% CI [.79-.82]), which outperformed participants in all three

conditions: No assistance ($D = 9.17, p < .001$), Prediction Only ($D = 5.68, p < .001$), and Prediction + Graphical Explanation ($D = 6.26, p < .001$).

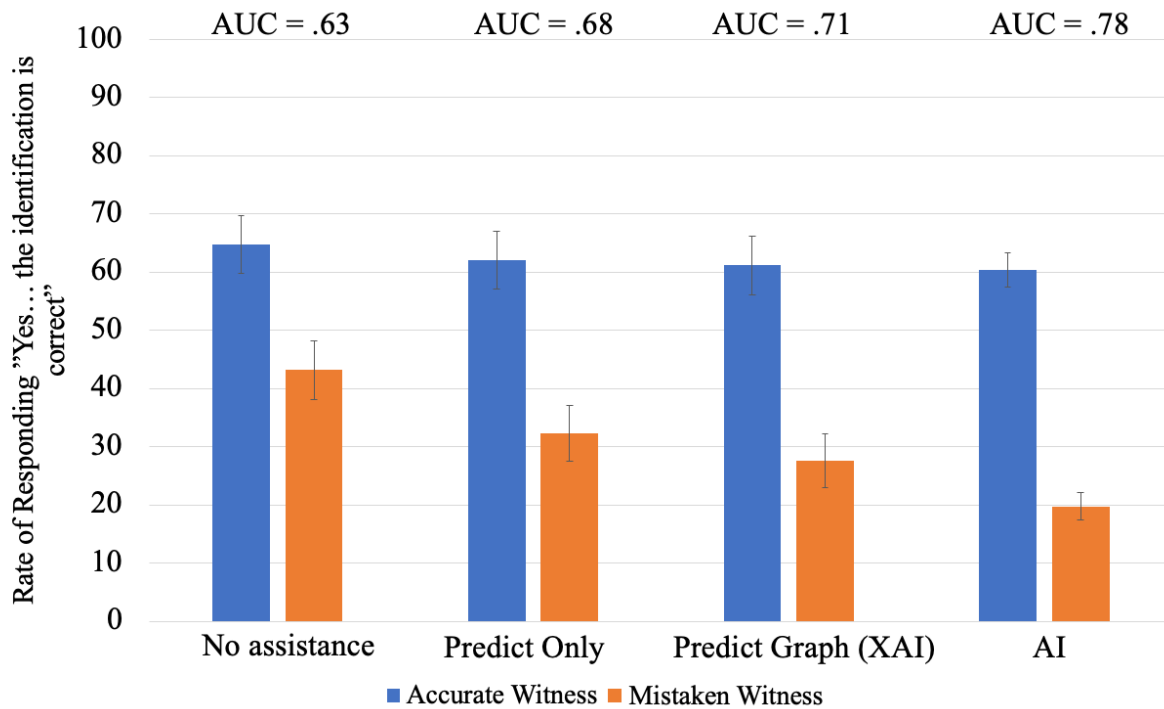


Figure 4. For recognition-based identifications, the rate of responding “Yes, I think the eyewitness’s identification is correct” to accurate and mistaken eyewitness identification trials in the different AI-assistance conditions. Errors bars indicate a 95% confidence interval.

Recognition-Based Identifications. Figure 4 shows that participants in the No-assistance condition discriminated between correct and incorrect recognition-based identifications roughly 63% of the time (AUC = .63; 95% CI [.59-.67])—a rate that is better than chance. AI-assistance in the form of the Prediction Only produced an AUC of .68 (95% CI [.64-.72]), which was not significantly different from the No-assistance condition, $D = 1.69, p = .126$. By contrast, the Prediction + Graphical Explanation condition did significantly improve discriminability (AUC = .71; 95% CI [.67-.74]) as compared to the No-assistance condition, $D = 2.78, p = .013$. Performance was comparable in the Prediction Only and Prediction + Graphical Explanation

conditions, $D = 1.08$, $p = .361$. Lastly, the AI alone (AUC = .78; 95% CI [.76-.80]) showed better discriminability than the No assistance ($D = 6.65$, $p < .001$), Prediction Only ($D = 4.65$, $p < .001$), and Prediction + Graphical Explanation ($D = 3.38$, $p = .002$) conditions.

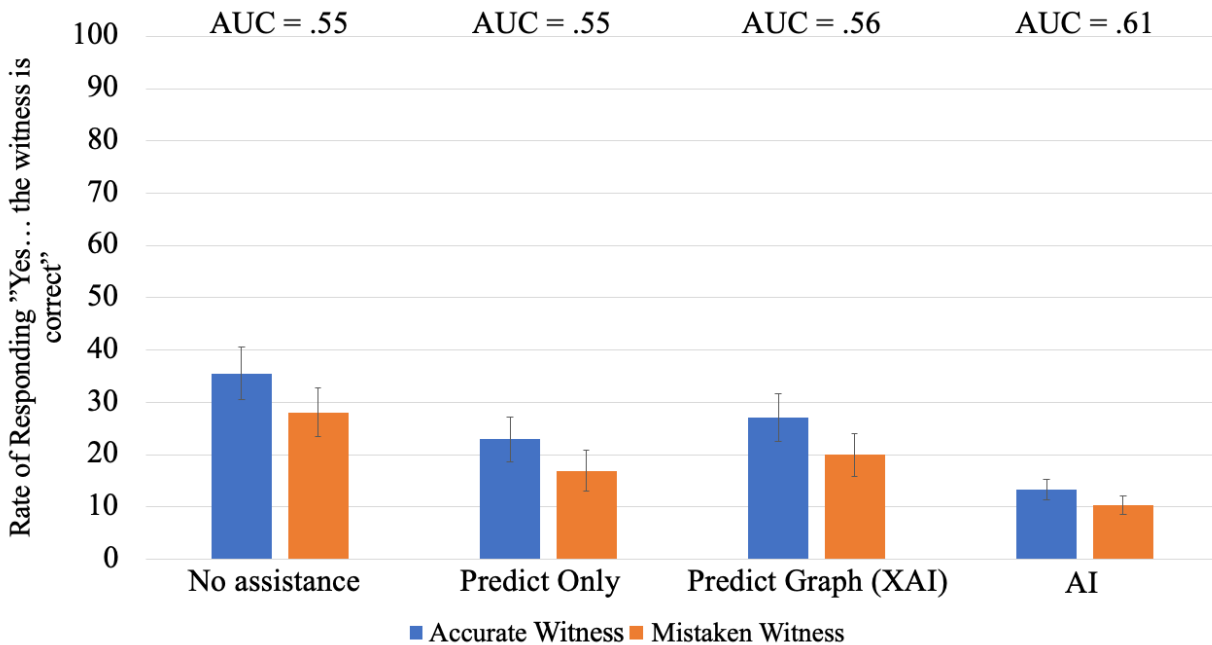


Figure 5. For familiarity-based identifications, the rate of responding “Yes, I think the eyewitness’s identification is correct” to accurate and mistaken eyewitness identification trials in the different AI-assistance conditions. Errors bars indicate a 95% confidence interval.

Familiarity-Based Identifications. Figure 5 shows the rates at which participants believed familiarity-based identifications were correct. Overall, discrimination was poor, especially compared to feature- and recognition-based identifications. In the No-assistance condition, participants’ discriminability was slightly above chance (AUC = .55; 95% CI [.51-.59]). However, AI-assistance did not improve performance. Discriminability was not significantly different in the No-assistance condition than in either the Prediction Only (AUC = .55; 95% CI [.51-.60]) condition, $D = 0.21$, $p = .879$, or the Prediction + Graphical Explanation (AUC = .56; 95% CI [.52-.60]) condition, $D = 0.44$, $p = .763$. Additionally, the Prediction Only and the

Prediction + Graphical Explanation conditions led to similar levels of performance, $D = 0.23$, $p = .837$. The AI alone (AUC = .61; 95% CI [.58-.63]) outperformed participants in the No-assistance condition ($D = 2.30$, $p = .045$), but performed similarly to participants in both AI-assistance conditions: Prediction Only ($D = 2.06$, $p = .061$), Prediction + Graphical Explanation ($D = 1.77$, $p = .120$).

We also examined how discriminability differed across the three types of identifications. Collapsed across the AI-assistance conditions, discriminability was better for feature-based (AUC = .65; 95% CI [.62-.67]) than familiarity-based (AUC = .55; 95% CI [.53-.58]) identifications, $D = 5.57$, $p < .001$, but was comparable to recognition-based identifications (AUC = .67; 95% CI [.65-.69]), $D = 1.41$, $p = .157$. Additionally, discriminability for recognition-based identifications was better than for familiarity-based identifications, $D = 7.01$, $p < .001$.

Overall, as compared to our No-assistance control condition, AI-assistance improved people's ability to distinguish between correct and incorrect identifications that are accompanied by recognition and featural justifications. However, participants in these conditions still performed worse than the AI alone. In contrast, discriminability was consistently poor for familiarity-based identifications, regardless of whether the participant received AI-assistance or not. Moreover, for familiarity-based identifications, participant performance was comparable to the AI's performance alone, with the exception of the no assistance condition. Finally, when collapsed across AI-assistance conditions, discriminability was worse for familiarity-based identifications than either feature- or recognition-based identifications.

Response Bias

We examined how statement-type and AI-assistance affected a participant's bias to believe that the witness's identification is correct. For each participant, we calculated a c-score (Macmillan & Creelman, 1990; Stanislaw & Todorov, 1999) using the *psycho* package (Makowski, 2018). The c-score provides an index of an individual's response bias, with an unbiased participant having a score of 0. More positive c-scores indicate a more skeptical bias (i.e., more likely to respond "No, I think the eyewitness is incorrect") and more negative c-scores indicate a more trusting bias (i.e., more likely to respond "Yes, I think the eyewitness is correct"). We ran a linear model using the *lme4* package (Bates et al., 2015) predicting the c-score from the factors of statement type and AI-assistance. In Wilkinson-Rogers (1973) notation the model was: $Cscore \sim StatementType * AIassistance$. Our model showed significant main effects of Statement Type $\chi^2(2) = 78.44, p < .001$ and AI-assistance $\chi^2(2) = 10.71, p < .001$. The interaction between Statement Type and AI-assistance was non-significant $\chi^2(4) = 2.17, p = .074$. In regard to Statement Type, participants showed a strong skeptical bias for familiarity-based identification ($M = 0.56; 95\% CI [0.51-0.61]$) but were relatively unbiased for identifications accompanied by featural ($M = -0.04; 95\% CI [-0.10-0.01]$) and recognition ($M = 0.04; 95\% CI [-0.02-0.09]$) statements. Focusing on AI-assistance condition, participants who received no AI-assistance showed a slight skeptical bias ($M = 0.08; 95\% CI [0.02-0.14]$), but this bias was stronger in both the Prediction Only ($M = 0.24; 95\% CI [0.17-0.29]$) and the Prediction + Graphical Explanation conditions ($M = 0.23; 95\% CI [0.18-0.30]$).

AI Usefulness Scores

We predicted that AI-assistance would more greatly influence discriminability on the part of participants who perceived the AI as more (vs. less) useful. To examine this prediction, we median split participants in each condition by their AI usefulness scores. We then constructed a ROC curve for those above and below the median in each condition. For participants who evaluated feature-based identifications, perceptions of AI usefulness did not significantly affect participant performance in the Prediction Only (high usefulness AUC = .68; 95% CI [.62-.73], low usefulness AUC = .68; 95% CI [.63-.73]; $D = 0.18, p = .860$) or the Prediction + Graphical Explanation condition (high usefulness AUC = .69; 95% CI [.64-.74], low usefulness AUC = .64; 95% CI [.58-.70]; $D = 1.22, p = .722$). For participants who evaluated identifications accompanied by recognition statements, discrimination ability was comparable for those who found the AI highly useful and those who did not—both in the Prediction Only (high usefulness AUC = .70; 95% CI [.64-.75], low usefulness AUC = .66; 95% CI [.60-.72]; $D = 0.89, p = .722$) and the Prediction + Graphical Explanation conditions (high usefulness AUC = .72; 95% CI [.66-.77], low usefulness AUC = .69; 95% CI [.64-.75]; $D = 0.55, p = .722$). Similarly, for participants who evaluated familiarity-based identifications, discriminability was no different for those who found the AI highly useful and those who did not. This pattern was present in both the Prediction Only (high usefulness AUC = .57; 95% CI [.51-.63], low usefulness AUC = .55; 95% CI [.49-.60]; $D = .55, p = 0.72$) and the Prediction + Graphical Explanation conditions (high usefulness AUC = .57; 95% CI [.52-.63], low usefulness AUC = .55; 95% CI [.49-.61]; $D = 0.56, p = .722$). Contrary to our prediction, we did not find evidence to support a difference in discrimination ability between those who found the AI more vs. less useful.

Discussion

Eyewitness misidentifications have contributed to a majority of wrongful convictions that were later overturned by DNA evidence (Innocence Project, 2023). Surprisingly, few studies have investigated people's ability to distinguish between accurate and inaccurate eyewitnesses, and no tool currently exists that can improve this ability. AI-assistance has the potential to assist police officers in interpreting eyewitness identifications and confidence statements.

Consistent with previous work (e.g., Smalarz & Wells, 2014) we find that, in the absence of AI-assistance, people show a modest ability to discriminate between correct and incorrect eyewitness identifications. But does AI-assistance improve discrimination? We tested this question by comparing participants who received no-assistance in evaluating mock-witnesses' statements to participants who received either an AI's prediction alone, or this prediction with an accompanying explanation (XAI).

AI-assistance generally improved people's ability to discriminate between eyewitness lineup identifications that were correct and incorrect when these identifications were accompanied by either an expression of recognition/recollection (e.g., "I remember him") or a featural justification (e.g., "His eyes stood out"). As compared to the No-assistance control condition, AI-assistance increased discriminability in all conditions, with the exception of a lack of improvement in the Prediction Only condition for recognition-based identifications. For familiarity-based identifications, however, AI-assistance did not improve people's ability to discriminate between correct and incorrect eyewitness identifications. A potential explanation for this lack of improvement in performance is that the AI's level of discriminability is very poor for familiarity-based identifications and much worse than it is for feature- or recognition-based

identifications. Consequently, there is less of an opportunity for an improvement in participants' discriminability when they receive assistance from the AI.

Consistent with our expectation, participants' discriminability was worse for familiarity-based identifications than it was for feature- or recognition-based identifications. A likely reason for this poorer discriminability is that the confidence-accuracy relationship for familiarity-based identifications is much weaker than for feature- or recognition-based identifications (e.g., Dobolyi & Dodson, 2018; Grabman et al., 2019). In other words, an eyewitness's level of confidence is less diagnostic of accuracy for familiarity-based identifications so, an evaluator must interpret a weaker signal of accuracy for these identifications than for feature-based or recognition-based identifications. Furthermore, these findings are in line with reality monitoring research (e.g., Johnson, Hashtroudi, & Lindsay, 1993) as people were better able to predict the accuracy of lineup identifications that were accompanied by specific featural or recollective details than identifications that were accompanied by feelings of familiarity.

We predicted that participants' perceptions of AI-usefulness would be an important moderating variable. Specifically, participants who found the AI more (vs. less) useful would show better discrimination. Contrary to our hypothesis, we found no evidence to support this prediction—discriminability was comparable for those who found the AI highly useful and those who did not.

Does the way in which we convey AI-assistance affect participants' discriminability? We observed comparable levels of discriminability when participants received AI-assistance in the form of either a graphical explanation (e.g., XAI) or simply showing the AI's prediction. This finding is consistent with previous studies that show that XAI provides little benefit as compared to simply presenting AI advice by itself (e.g., Alufaisan et al., 2020). However, the lack of added

benefit in the current study of a graphical explanation may be due to the relatively impoverished set of predictors of eyewitness accuracy (i.e., only the AI's weighting of individual words).

Previous studies that have used XAI have included a larger set of predictors. For example, Liu et al. (2021) used a prediction task in which participants had to decide whether a person would or would not violate the terms of their pretrial release. In their XAI condition, participants were provided with multiple factors (e.g., age, race, prior arrests, etc.) to consider and whether each factor predicted the person would or would not violate their terms. So, showing participants more predictors of the accuracy of the eyewitness's identification (e.g., decision-time, age, etc.) may improve discriminability in our XAI condition.

In sum, for the first time, we show that AI-assistance can improve a person's ability to discriminate between correct and incorrect eyewitness identifications. For identifications accompanied by recognition and featural justifications, participants who received AI-assistance showed improved discriminability compared to those who received no assistance. For familiarity-based identifications, however, discrimination ability was comparable across all three AI-assistance conditions. This study is an important step in investigating the use of AI-assistance in the eyewitness lineup identification context.

Author Contributions

C.S. Dodson developed the study concept. All authors contributed to the study design. L. E.

Kelso performed primary data collection, analysis, interpretation, and wrote the first draft. C. S.

Dodson, J. H. Grabman, and D. G. Dobolyi provided critical revisions. All authors approved the

final version of the manuscript.

References

- Abbasi, A., Dobolyi, D., Vance, A., & Zahedi, F. M. (2021). The Phishing Funnel Model: A Design Artifact to Predict User Susceptibility to Phishing Websites. *Information Systems Research, 32*(2), 410-436. <https://doi.org/10.1287/isre.2020.0973>
- Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does Explainable Artificial Intelligence Improve Human Decision Making? In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, 6618-6626. <https://doi.org/10.1609/aaai.v35i8.16819>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld. D. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-16. <https://doi.org/10.1145/3411764.3445717>
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Beaudry, J. L., Lindsay, R. C. L., Leach, A.M., Mansour, J. K., Bertrand, M. I., & Kalmet, N. (2015). The effect of evidence type, identification accuracy, line-up presentation, and line-up administration on observers' perceptions of eyewitnesses. *Legal and Criminological Psychology, 20*(2), 343-364. <https://doi.org/10.1111/lcrp.12030>
- Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior, 29*, 279-301.

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1), 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2021). Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 401-413).
- Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 454-464. <https://doi.org/10.1145/3377325.3377498>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5, 1-21. <https://doi.org/10.1145/3449287>
- Clark-Foos, A., Brewer, G., & Marsh, R. L. (2015) Judging the reality of others' memories. *Memory*, 23(3), 427-436. <http://dx.doi.org/10.1080/09658211.2014.893364>.
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristics curves: a nonparametric approach. *Biometrics*, 44(3), 837-845. <https://doi.org/10.2307/2531595>.
- Dobbins, I. G., & Kantner, J. (2019). The language of accurate recognition memory. *Cognition*, 192, 103988. <https://doi.org/10.1016/j.cognition.2019.05.025>.
- Dobolyi, D. G., & Dodson, C. S. (2018). Actual vs. Perceived Eyewitness Accuracy and

- Confidence and the Featural Justification Effect. *Journal of Experimental Psychology: Applied*, 24(4), 543-563. <http://dx.doi.org/10.1037/xap0000182>
- Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting Eyewitness Expressions of Confidence: The Featural Justification Effect. *Law and Human Behavior*, 39(3), 266-280. <https://doi.org/10.1037/lhb0000120>
- Egelman, S., Cranor, L. F., & Hong, J. (2008). You've been warned: An empirical study of the effectiveness of web browser phishing warnings. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1065-1074. <https://doi.org/10.1145/1357054.1357219>
- Fitzgerald, R. J., Rubínová, E., & Juncu, S. (2021). Eyewitness identification around the world. In A. M. Smith, M. P. Togli, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks*. Taylor & Francis. <https://doi.org/10.17605/OSF.IO/KN6R5>.
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI. *Management Information Systems Quarterly*, 45, 1527-1556.
- Gamoran, A., Lieberman, L., Gilead, M., Dobbins, I. G., & Sadeh, T. (2024). Detecting recollection: Human evaluators can successfully assess the veracity of others' memories. *Proceedings of the National Academy of Sciences*, 121(22), e2310979121. <https://doi.org/10.1073/pnas.2310979121>.
- Grabman, J. H., Dobbins, I. G., & Dodson, C. S. (2024). Comparing human evaluations of eyewitness statements to a machine learning classifier under pristine and suboptimal

lineup administration procedures. *Cognition*, 251, 105876.

<https://doi.org/10.1016/j.cognition.2024.105876>

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting High Confidence Errors in Eyewitness Memory: The Role of Face Recognition Ability, Decision-Time, and Justifications. *Journal of Applied Research in Memory and Cognition*, 8(2), 233-243. <https://doi.org/10.1016/j.jarmac.2019.02.002>

Grabman, J. H., & Dodson, C. S. (2024). Unskilled, underperforming, or unaware? Testing three accounts of individual differences in metacognitive monitoring. *Cognition*, 242, 105659.

Innocence Project (2023). Eyewitness misidentification. *Innocence Project*.

<https://innocenceproject.org/eyewitness-misidentification/>.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin*, 114(1), 3-28.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, 88(1), 67-85.

<https://doi.org/10.1037/0033-295X.88.1.67>.

Kaminski, K. S., & Sporer, S. L. (2017) Discrimination Between Correct and Incorrect Eyewitness Identifications: The Use of Appropriate Cues. *Journal of Experimental Psychology: Applied*, 23(1), 59-70. <http://dx.doi.org/10.1037/xap0000110>.

Kelso, L. E., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (2024). Does AI-assistance mitigate biased evaluations of eyewitness identifications? *Journal of Applied Research in Memory and Cognition*, X(X), XX-XX. <https://doi.org/10.1037/mac0000192>.

Kelso, L. E., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (2024). Can AI-assistance Improve a Person's Ability to Discriminate Between Correct and Incorrect Eyewitness Identifications? <https://doi.org/10.17605/OSF.IO/MCFZ9>

- Lemon, J. (2006). Plotrix: a package in the red light district of R. *R-News*, 64(4), 8-12.
- Macmillan, A. N., & Creelman, C. D. (1990). Response Bias: Characteristics of Detection Theory, Threshold Theory, and “Nonparametric” Indexes. *Psychological Bulletin*, 107(3), 401-413. <https://doi.org/10.1037/0033-2909.107.3.401>
- Makowski, D. (2018). The Psycho Package: An Efficient and Publishing-Oriented Workflow for Psychological Science. *Journal of Open Source Software*, 3(33), 470. <https://doi.org/10.21105/joss.00470>.
- National Registry of Exonerations. (2023). *Annual report 2023*. <https://www.law.umich.edu/special/exoneration/Documents/2023%20Annual%20Report.pdf>
- R Core Team. (2024). *A language and environment for statistical computing*. R Foundation for Statistical Computing (Version 4.3.3) [Computer software]. <https://www.R-project.org/>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 1-8. <https://doi.org/10.1186/1471-2105-12-77>.
- Schemmer, M., Hemmer, P., Nitsche, M., Köhl, N., & Vössing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAI/ACM Conference on AI, Ethics, and Society*, 617-626. <https://doi.org/10.1145/3514094.3534128>
- Schooler, J. W., Gerhard, D., & Loftus, E. F. (1986). Qualities of the Unreal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(2), 171-181.
- Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2022). The language of accurate and

- inaccurate eyewitnesses. *Journal of Experimental Psychology: General*, 151(6), 1283-1305. <https://doi.org/10.1037/xge0001152>
- Slane, C. R., & Dodson, C. S. (2022). Eyewitness Confidence and Mock Juror Decisions of Guilt: A Meta-Analytic Review. *Law and Human Behavior*, 46(1), 45-66. <https://doi.org/10.1037/lhb0000481>.
- Smalarz, L., & Wells, G. L. (2014). Post-Identification Feedback to Eyewitnesses Impairs Evaluators' Abilities to Discriminate Between Accuracy and Mistaken Testimony. *Law and Human Behavior*, 38(2), 194-202. <https://doi.org/10.1037/lhb0000067>. <https://doi.org/10.3758/BF03207704>
- Smalarz, L., Yang, Y., & Wells, G. L. (2021). Eyewitnesses' free-report verbal confidence statements are diagnostic of accuracy. *Law and Human Behavior*, 45(2), 138-151. <https://doi.org/10.1037/lhb0000444>
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137-149.
- Stebly, N. K., & Wells, G. L. (2023) In their own words: Verbalizations of real eyewitnesses during identification lineups. *Psychology, Public Policy, and Law*, 29(3), 272-287. <https://doi.org/10.1037/law0000386>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3), 425-278. <https://doi.org/10.2307/30036540>
- Wang, X., Lu, Z., & Yin, M. (2022). Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM web conference 2022* (pp. 1697-1708). <https://doi.org/10.1145/3485447.3512240>.

- Wellcome Library, London. (2018). *L0059158 Eight Ishihara charts for testing colour blindness, Europe* [Photograph]. Retrieved from https://commons.wikimedia.org/wiki/File:Eight_Ishihara_charts_for_testing_colour_blindness,_Europe_Wellcome_L0059158.jpg
- Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22, 392-399. <https://doi.org/10.2307/2346786>
- Yates, S.Q. (2017). Memorandum for Heads of Department Law Enforcement Components All Department Prosecutors. U.S. Department of Justice, Office of the Deputy Attorney General, 1-12. Retrieved from <https://www.justice.gov/file/923201/download>
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp.295-305). <https://doi.org/10.1145/3351095.3372852>.