"even explanations will not help in trusting [this] fundamentally biased system" A Predictive Policing Case-Study

Mehrotra, Siddharth; Gadiraju, Ujwal; Bittner, Eva; Van Delden, Folkert; M. Jonker, Catholijn; Tielman, Myrthe L.

DOI
10.1145/3699682.3728343

Licence
CC BY-NC

Publication date
2025

Document Version
Final published version

Published in
UMAP 2025 - Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization

Citation (APA)
Mehrotra, S., Gadiraju, U., Bittner, E., Van Delden, F., M. Jonker, C., & Tielman, M. L. (2025). "even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study. In *UMAP 2025 - Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 51-62). (UMAP 2025 - Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization). Association for Computing Machinery (ACM). https://doi.org/10.1145/3699682.3728343

Latest updates: https://dl.acm.org/doi/10.1145/3699682.3728343

RESEARCH-ARTICLE

“Even explanations will not help in trusting [this] fundamentally biased system”: A Predictive Policing Case-Study

SIDDHARTH MEHROTRA, Delft University of Technology, Delft, Zuid-Holland, Netherlands

Citizen trust in Public Sector AI Systems | Human-AI Interaction | AI Safety

UJWAL GADIRAJU, Delft University of Technology, Delft, Zuid-Holland, Netherlands

EVA BITTNER, University of Hamburg, Hamburg, Hamburg, Germany

FOLKERT VAN DELDEN, Delft University of Technology, Delft, Zuid-Holland, Netherlands

CATHOLIJN M JONKER, Leiden University, Leiden, Zuid-Holland, Netherlands

MYRTHE LOTTE TIELMAN, Delft University of Technology, Delft, Zuid-Holland, Netherlands

Open Access Support provided by:

Delft University of Technology

University of Hamburg

Leiden University

PDF Download
3699682.3728343.pdf
16 January 2026
Total Citations: 0
Total Downloads: 1222

Published: 16 June 2025

Citation in BibTeX format

UMAP '25: 33rd ACM Conference on User Modeling, Adaptation and Personalization
June 16 - 19, 2025
New York City, USA

Conference Sponsors:
SIGCHI

UMAP '25: Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (June 2025)
https://doi.org/10.1145/3699682.3728343
ISBN: 9798400713132

“Even explanations will not help in trusting [this] fundamentally biased system”: A Predictive Policing Case-Study

Siddharth Mehrotra
TU Delft
Delft, Netherlands
s.mehrotra@tudelft.nl

Ujwal Gadiraju
Web Information Systems, Software Technology Department
TU Delft
Delft, Netherlands
u.k.gadiraju@tudelft.nl

Eva Bittner
University of Hamburg
Hamburg, Germany
bittner@informatik.unihamburg.de

Folkert van Delden
TU Delft
Delft, Netherlands
F.E.vanDelden@tudelft.nl

Catholijn M. Jonker
TU Delft
Delft, Netherlands
Leiden University
Leiden, Netherlands
c.m.jonker@tudelft.nl

Myrthe L. Tielman
TU Delft
Delft, Netherlands
m.l.tielman@tudelft.nl

Abstract

In today’s society, where Artificial Intelligence (AI) has gained a vital role, concerns regarding user’s trust have garnered significant attention. The use of AI systems in high-risk domains have often led users to either under-trust it, potentially causing inadequate reliance or over-trust it, resulting in over-compliance. Therefore, users must maintain an appropriate level of trust. Past research has indicated that explanations provided by AI systems can enhance user understanding of when to trust or not trust the system. However, the utility of presentation of different explanations forms still remains to be explored especially in high-risk domains. Therefore, this study explores the impact of different explanation types (text, visual, and hybrid) and user expertise (retired police officers and lay users) on establishing appropriate trust in AI-based predictive policing. While we observed that the hybrid form of explanations increased the subjective trust in AI for expert users, it did not led to better decision-making. Furthermore, no form of explanations helped build appropriate trust. The findings of our study emphasize the importance of re-evaluating the use of explanations to build [appropriate] trust in AI based systems especially when the system’s use is questionable. Finally, we synthesize potential challenges and policy recommendations based on our results to design for appropriate trust in high-risk based AI-based systems.

CCS Concepts

• **Computing methodologies → Artificial intelligence**; • **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Appropriate Trust, Explainable AI, Predictive Policing, Trust

UMAP ’25, New York City, NY, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1313-2/25/06
https://doi.org/10.1145/3699682.3728343

ACM Reference Format:
Siddharth Mehrotra, Ujwal Gadiraju, Eva Bittner, Folkert van Delden, Catholijn M. Jonker, and Myrthe L. Tielman. 2025. “Even explanations will not help in trusting [this] fundamentally biased system”: A Predictive Policing Case-Study. In *33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’25), June 16–19, 2025, New York City, NY, USA*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3699682.3728343

1 Introduction

Artificial Intelligence (AI) is rapidly reshaping public organizations globally, mainly through machine learning approaches that automate routine administrative tasks and support decision-making [11]. One of the key components to achieving effective decision-making is a user’s appropriate trust in AI systems. Despite recent efforts towards enhancing trust in algorithmic decision-making systems (e.g., adding *explanations* [20, 90], *human oversight* [71, 81, 85], and *confidence scores* [6, 91]), comparatively little attention has been paid to building appropriate trust in them. Both under-trust and over-trust are deemed inappropriate [45, 72, 76], instead we require trust to be appropriate. Under-trust can lead to under-reliance, and over-trust can lead to over-compliance, which can negatively impact the task. Therefore, in this work we study whether we can improve the appropriateness of trust in an AI decision support system (*goal*). We propose to do this through explanations (*means*), and position our work in the context of AI-based predictive policing (*context*) as a high-risk domain.

We choose AI-based predictive policing as our use case primarily because it represents a critical domain where appropriate trust is pivotal due to the high-stakes nature of decision-making. Additionally, the use of AI in predictive policing has been the subject of extensive debate for years, with numerous studies highlighting biases inherent in data collection practices and their impact on AI-driven decisions [3, 26, 47, 51, 62, 63]. Despite these concerns, predictive policing systems continue to be used globally [67]. Therefore, in this study we explore whether providing explanations can help users critically evaluate AI decision-making in this use-case, encouraging introspection and ultimately NOT promoting trust but fostering appropriate trust.

We draw inspiration from the ASPECT model by Jameson et al. [35], as a useful means to identify the different facets of our complex use case. The ASPECT acronym refers to the first letters of six patterns: Attributes, Social influence, Policies, Experience, Consequences, and Trial and error. Each pattern may be perceived as an "aspect" of design choices, denoting a particular perspective or approach to addressing a problem or research question. Aligned with the ASPECT model, our approach with this use-case is based on attributes (explanations), social (socially relevant topic), policy (policy level implications), experience (user expertise), consequence (effect of explanation), and trial & error (human-AI agreement) based choice patterns of the ASPECT model.

While AI systems can provide explanations in multiple formats, selecting the optimal design remains challenging. These formats include visual explanations like saliency maps [1], textual explanations using words and phrases, and analytical explanations that enable users to explore both data and model [32, 40]. Although each method has garnered significant attention independently, comparative studies examining their relative effectiveness across different contexts and user groups remain limited [68, 73, 75, 84]. Notably, there is a particular gap in understanding how these different presentation methods impact appropriate trust in high-risk domains. Our research addresses this gap by examining user interaction and perception across textual, visual, and hybrid explanation formats to determine their effectiveness in fostering appropriate trust.

According to Ribera & Lapedriza [74], the goal of XAI in any context depends not only on the presentation of explanations but also on the type of end-user that is on the receiving end of the explanations. For example, previous research has shown that the hybrid form of explanations significantly improves correct understanding for lay users compared to visual explanations [84]. However, little work has been done so far to compare the utility of different explanation methods in building appropriate trust between expert users and lay users [43]. In the predictive policing use-case, this comparison is relevant as some police officers might have less professional experience to draw on than others, for example, police officers who have recently joined the department. Therefore, in this study, we compared the utility of different types of explanations with both expert and lay users (*moderation factor*) linked with the experience pattern of the ASPECT model.

To better understand the moderating factor of user expertise in our study, we performed an application-grounded evaluation followed by a human-grounded evaluation. Human-grounded evaluation is appealing when orchestrating experiments with the target user groups is challenging [39, 42]. This is in line with Doshi-Velez & Kim [21] who articulate and argue for the value of different layers of evaluation within XAI focusing on functionality, grounding, and presentation. Our sample of police officers corresponds to application-grounded evaluation which includes experts and real tasks. On the other hand, our sample of lay users corresponds to human-grounded evaluation which includes real humans performing identical tasks. Our contribution in this work is centered around a specific use-case that warrants assessment within the framework of application and human-grounded evaluation.

We aim to address the following research questions:

RQ1: What effect do different types of explanations (no explanation, textual, visual or hybrid) have on building appropriate trust in AI-based predictive policing systems?
RQ2: How does human trust in the AI assistant change given these different types of explanations?
RQ3: Do lay users or experts find these explanations useful in making a decision?
RQ4: Is there a moderating effect of user expertise influence the role of explanations in establishing appropriate or subjective trust?

We investigated the first question by prompting users with a selection of hotspots for predictive policing that gauge their understanding of the explanation at hand and see whether they can appropriately trust the system. For measuring appropriate trust, we adopt several existing measures from the literature. The second question is addressed by asking participants to rate their perceived trust in the AI assistant, and the third by rating the usefulness of the explanations over multiple rounds in our study. Finally, we address the last question by comparing participants with different expertise (police officers & lay users) and studying the role of explanations in building appropriate and subjective trust.

Original Contributions. Through our work in this paper, we make the following contributions:

1. We present the first study exploring the effect of different form of explanations on building appropriate trust using a prototypical system in the context of AI-based predictive policing.
2. We illustrate the effect of user expertise on different form of explanations for building appropriate trust.
3. By conducting two user studies (*N=192, 12 experts and 180 layusers*), we show that even with different form of explanations, participants often end up in the trap of confirmation biases and no form of explanation helped in fostering appropriate trust.
4. Based on our results, we highlight research challenges and recommendations for the design of public sector AI systems.

What this work is <u>not</u> about?[1]

1. Ethical implications of predictive policing and people affected by such systems.
2. Target users of the AI-based predictive policing system.

2 Background and Related Work

2.1 Appropriate Trust

Appropriate Trust in AI systems has rapidly become an important area of focus for researchers and practitioners. As technology evolved from automated machines to decision aids, virtual avatars, robots, and AI teammates, appropriate trust has been studied in depth and breadth across various domains.

2.1.1 Definitions. It is important to understand how we define appropriate human trust in AI (Human-AI trust) when trying to achieve it. On the one hand, the definitions of appropriate trust are linked to system performance or reliability, such as by McBride and

[1] The decision to exclude discussions on the ethical implications of predictive policing and the individuals affected by such systems is not one made lightly, but rather stems from the specific focus of our research. While acknowledging the profound ethical considerations surrounding predictive policing, our study is primarily concerned with investigating the role of different forms of explanations in building appropriate trust.

Morgan [54], McGuirl and Sarter [56], Ososky et al. [71], Yang et al. [90]. On the other hand, the definitions of appropriate trust are related to trustworthiness and beliefs such as by Danks [16], Ferreira Gomes Centeio Jorge et al. [27], Mehrotra et al. [61]. Inevitably, despite the crucial role of appropriate trust in ensuring the successful use of AI systems, there is currently a fragmented overview of its understanding [48]. This conclusion resonates with [34] overview, who calls for definitions to be precisely defined and differentiated.

2.1.2 Use of Explanations for Fostering Appropriate Trust. A common method to achieve appropriate trust is by adding transparency to the system through explanations [44, 46, 61, 87]. Intuitively, this makes sense as understanding an AI system's inner workings and decision-making should, in theory, also allow a user to understand better when to trust or not trust a system to perform a task [19]. For example, it has been shown that an AI agent who displays its integrity in the form of explanations by being explicit about potential biases in data or algorithms achieved appropriate trust more often than being honest about capability or transparent about the decision-making process [61]. Similarly, Yang et al. results indicated that visual explanation led to users' appropriate trust in machine learning and improved appropriate use of the recommendations from the classifier [90].

On the flip side of the coin, some empirical evidence suggests that even technically sound AI explanations can result in harmful over-trust and over-reliance [5, 22, 37, 58, 91]. For example, studying about 'cognitive forcing functions' Buçinca et al. have shown that explanations with these functions were effective in trust calibration, as here the AI system adjusts to the user's attitude and behaviour following the signs of over- and under-trust [10]. But the study, in contrast, highlights that people do not cognitively engage with explanations. Similarly, Bertrand et al. find that providing feature-based explanations does not improve appropriate reliance or understanding compared to not providing any explanation [8]. Therefore, given this clear lack of consensus more work into the effect of XAI on appropriate trust is warranted [48].

Other research in XAI has explored how expertise influences the perception of explanations in building trust. For instance, Simkute et al. [82] highlight the importance of tailoring explanations to account for differences in reasoning between experts and lay users. Naturally, experts tend to critique explanations more rigorously, which can sometimes lead to insufficient trust, whereas lay users are more prone to over-reliance on the system [7, 77]. A common denominator highlights that explanations must support either trust building for experts, or critical thinking for lay users.

2.2 AI-based Predictive Policing & Trust

In light of the EU AI Act's regulatory framework, the use of predictive policing systems, which heavily relies on advanced data analysis methods, may come under scrutiny and be subject to compliance with the Act's provisions on high-risk AI applications in law enforcement [15]. This introduces a dimension of accountability and transparency in deploying predictive policing technologies within the legal and ethical parameters [3, 31].

Currently, there are four main applications of predictive policing being used in European and American police departments: CAS (Crime Anticipation System) in the Netherlands[2], ProMap and PredPol in UK[3], and Soundthinking in the US[4]. Multiple studies have been conducted to highlight the issues with these applications [29, 52, 70]. For example, Meijer et al.'s study highlights two patterns of algorithmization of government bureaucracy - the 'algorithmic cage' (Berlin, more hierarchical control) and the 'algorithmic colleague' (Amsterdam, room for professional judgment) [62]. Specifically looking at trust, a study by Selten et al. shows that police officers trust and follow AI recommendations congruent with their intuitive professional judgment [80].

Studies examining predictive policing systems consistently reveal systemic limitations and biases, particularly regarding the over-policing of marginalized communities [2, 24, 49]. Lum & Isaac's research demonstrates how these systems amplify historical biases, intensifying surveillance in already over-policed areas while simultaneously reducing algorithmic accountability [49]. This issue is compounded by the self-reinforcing feedback loops identified by Ensign et al., where police resources are repeatedly directed to the same neighborhoods independent of actual crime rates [24]. Overall, these findings underscore the critical importance of appropriate trust in predictive policing systems, making this domain an essential case study for examining how explanations can foster appropriate trust.

3 Study Design

In our main user study, we sought to understand the effect of different explanations on appropriate trust and the role of user expertise. To understand the role of user expertise, we conducted two user studies. In the first user study, we recruited 12 ex-police officers from the Dutch police who retired in the last five years. We refer to this group of participants as "*Expert users*". In the second user study, we recruited 180 crowdsourced workers without experience with predictive policing systems. We refer to this second group of participants as "*Lay users*". As indicated in section 1, our choice of recruiting two different sets of users is necessary to answer RQ3 & RQ4 and is situated in line with prior work of [21]. Finally, before data collection we preregistered our design and data analysis plan[5]. With our study design, our aim was to closely replicate the decision-making contexts of the real world police teams by closely simulating the real world conditions.

3.1 Designing AI system's Explanations

We conducted a preliminary study with three expert users to inform us about the design of explanations (details available in footnote 5). Building on insights from this preliminary study, we sought to design our explanations. Also, we added weather and escape route information to the explanation based on an insight which emerged during our preliminary study. Once we decided on the content of the explanations, we looked at established guidelines in the literature on crafting them. We selected the guidelines by Szymanski et al. [84] because they conducted a state-of-the-art literature survey and a formative study on XAI. The guidelines

[2] https://kombijde.politie.nl/vakgebieden/ict/predictiv

[3] https://www.lawsociety.org.uk/topics/research/algorithm-use-in-the-criminal-justice-system-report

[4] https://www.soundthinking.com/law-enforcement/crime-analysis-crimetracer/

[5] https://osf.io/hka58/?view_only=c18a9092440f4509a9225add1af51f91

include (a) quantifying each parameter's contribution to prediction, (b) what parameters lead to predictions, (c) instances with similar predictions, (d) locating regions about uncertainity, and (e) displaying an overall predictions. To explain the predictive policing system's decision in a textual form, we generated sentences per input parameter using the template described by Hohman et al. [33]: *The system predicts a higher likelihood of incidents in hotspot A/ B/ C/ D based on (historical crime data OR proximity to dense forest/ highway/ sea OR last arrest of offenders)* ***[A]****. The X% confidence score reflects (strong/ weak) support* ***[B]****, and the remaining X% acknowledges potential unknowns like X* ***[C]****. Major contributing factors to this decision include C1 (X%), C2 (X%), and C3 (X%)* ***[D]****. Furthermore, a similar case was found in X's police records, where offenders were caught near X* ***[E]****. A strong/ no correlation with severe weather (snow or thunderstorms) was found while making this decision* ***[D]****. (Tip: Weather prediction for the next three days is X; allocate resources accordingly.*

Here, **[A]** denotes overall prediction, **[B]** denotes the confidence in the prediction, and **[C]** shows regions where the model prediction was uncertain. **[D]** quantifies each parameter's contributions, and the name of the parameters, and **[E]** denotes instances that have similar predictions. The contributing factors were among the following: (C1) Historical crime data, (C2) Geographical information, (C3) Time and day of the week, (C4) Weather information, (C4) Demographic statistics, (C5) Resource availability, and (C6) Socioeconomic data. To enable a fair comparison, the visual explanations contained the same information as the textual explanations. Figure 1 shows an example of a visual explanation used in the study. Finally, as prior research on designing hybrid explanations is limited [66], we based our design only on previous work done by [84], who combined visual explanations with text.

Traditional Investigation Notes: Based on our preliminary study, police officers often follow traditional methods (diary notes, intel from other units and department instructions) for investigation in conjunction with predictive policing systems. On the one hand these notes serves as the ecological validity (in real life police officers often use their diary notes for investigation) for the task and on the other hand they make sure that there is a 'joint' knowledge for our both groups of participants. An example of a note used in this study is as follows:

You have (less than a year OR more than three years of experience) in this area shown on the map ***[A]****. According to your diary notes, under the cover of darkness, the past offenders often slip through the labyrinth of narrow alleyways matching with the hotspot A/ B/ C/ D* ***[B]****. According to the intelligence department of the Police, the last fugitive vanished into the dense forest after following the alleyways* ***[C]****.* Here, **[A]** denotes the overall experience of the police officer in the selected area, **[B]** denotes diary notes and probable hotspot selection, and **[C]** shows the intel from the intelligence department.

4 First User Study - "Expert Users"

4.1 Participants

For our expert study, we recruited 12 retired police officers (aged between 65 and 70, 10M:2F) who retired in the last five years from the Dutch police. Our goal was to recruit experts who had prior experience with predictive policing systems or were in-charge of making decisions related to crime prevention. The retired police officers fulfilled these criteria as they were mostly in higher positions in their hierarchy before they retied. Furthermore, given the discussions around the new EU AI Act, police officers in the current force were hesitant to join our user study. Therefore, we decided to recruit retired police officers as they fit our goal of expert users. This study was approved by the Human Research Ethics Review Board of our university and was conducted in the Dutch language.

4.2 Methodology

4.2.1 Independent variable. **Explanations** (*categorical, between-subjects*). We assigned each participant to one of four configurations: (1) **No explanation**: participants saw hotspot selection by the AI assistant but not how this decision was made. (2) **Text-based Explanations**: participants saw the hotspot selection by the AI assistant and how this decision was made in textual form. (3) **Visual Explanations**: participants saw the hotspot selection by the AI assistant and how this decision was made in visual form. (4) **Hybrid (Text+Visual) Explanations**: participants saw a combination of text-based and visual explanation.

4.2.2 Dependent variables. (a) **Appropriate Trust** (*continuous*). Adapted from [41, 61, 90, 91]. These measures are described in section 4.2.4. (b) **Subjective Trust** (*continuous*). We used a self-reported global trust meter that captures changes in trust for each round ranged from completely distrust (-100) to completely trust (+100), adapted from [38, 61, 90]. (c) **Usefulness of Explanations** (*continuous*). The usefulness of explanations was measured on a 7-point Likert scale from Not at all helpful (1) to Very helpful (7), adapted from [90].

4.2.3 Descriptive and exploratory measurements. We use these variables to describe our sample and for exploratory analyses, but we do not conduct any conclusive hypothesis tests on them.

(a) **Age group** (*categorical*). Participants will select their age group from multiple choices. (b) **Level of education** (*categorical*). Participants will select the highest level of education they have completed. (c) **AI literacy** (*continuous*). Average score of the four items defined by [79]. (d) **Propensity to Trust** (*continuous*). Propensity to Trust scale by Merritt et al. [64] adapted to understand apriori trust in the predictive policing systems. (e) **Personal experiences as a police officer and use of AI** (1)) Have you ever worked with predictive policing systems in the past? and (2)) Do you have prior experience with the use of AI in predictive policing systems? (f) **Task stakes perception** (*continuous*). In this study we have considered scenario such as pick-pocketing as non-violent crime and sexual-offense as a violent crime based on the Dutch WODC Magazine Recidivism[6]. Since the stakes involved in a decision are subjective [36], we will capture task stakes perceptions using [50]. (g) **AI Confidence Score** (*categorical, within-subjects*) AI accuracy was communicated to participants as a part of the explanations. (1) High (*Confidence Score* $>$ *75%*) and (2) Low (*Confidence Score* $<$ *75%*). (h) **Geographical Experience** (*categorical, within-subjects*): Prior experience with policing about the shown geographic area on the map was communicated in the dairy notes. (1) Amount of

[6] https://magazines.wodc.nl/wodcmagazine/2019/03/high-impact-crime-hic

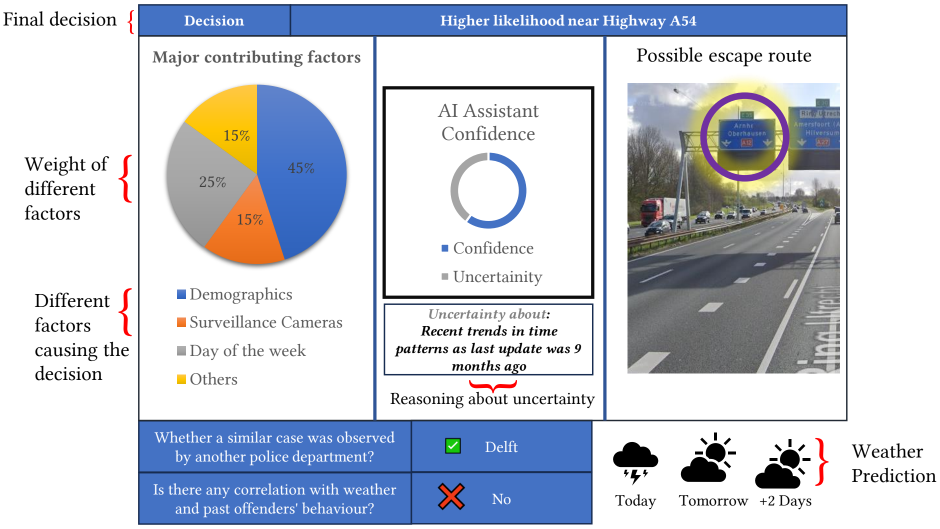

Final decision
Decision
Higher likelihood near Highway A54
Weight of different factors
Different factors causing the decision
Major contributing factors
15%
45%
25%
15%
■ Demographics
■ Surveillance Cameras
■ Day of the week
■ Others
AI Assistant Confidence
■ Confidence
■ Uncertainity
Uncertainty about:
Recent trends in time patterns as last update was 9 months ago
Reasoning about uncertainty
Possible escape route
Whether a similar case was observed by another police department?
☑ Delft
Is there any correlation with weather and past offenders' behaviour?
☒ No
Today
Tomorrow
+2 Days
Weather Prediction

Figure 1: Visual explanations for a selected instance from the user study.

professional experience : Limited (> *3 years experience*) and Amount of professional experience : High (< *3 years experience*).

4.2.4 Measurement of appropriate trust. In this study, we used two measurements of appropriate trust and calibrated trust each from prior research, see [41, 61, 90, 91]. We used distinctive measures for appropriate and calibrated trust based on the definitions provided in the literature [18, 69, 89]. For example, Mehrotra et al. show that different definitions and measures of appropriate and calibrated trust exist in the literature [59]. We argue that it is necessary to study multiple measures to understand when trust can be classified as appropriate or calibrated, as different measures may result in slightly different outcomes. Also, it is important to differentiate between trust, trustworthiness and reliance as illustrated by Tolmeijer et al [85]. In this paper, we therefore define trust as the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [45], trustworthiness as an antecedent of trust [53] and reliance as "a discrete process of engaging or disengaging" [45] with the AI system. Our measures are:

Measure 1 (App1): Appropriate trust is to [not] follow an [in]correct recommendation [90][7].

Measure 2 (App2): Appropriate trust occurs when (a) the human estimates that the AI agent is better at the task than the human, (b) also the actual Trustworthiness[8] (TW) of the AI agent is equal to or higher than the human's TW and (c) the human selects the AI agent for the task and *vice-versa* [60].

Measure 3 (Calib1): Switch percentage, the percentage of trials in which the participant switched from their initial prediction to use the AI's prediction as their final prediction [91].

Measure 4 (Calib2): Agreement percentage, the percentage of trials in which the participant's final prediction agreed with the AI's prediction [41].

Table 1: Categorization of measures of appropriate trust based on the work by Mehrotra et al. [61]

Round	Human TW	AI TW	AI	Human
1	High	High	Correct	Correct
2	Low	High	Correct	Incorrect
3	Low	Low	Correct	Incorrect
4	High	High	Incorrect	Correct
5	High	Low	Correct	Correct
6	Low	High	Incorrect	Incorrect
7	Low	Low	Incorrect	Correct
8	High	Low	Incorrect	Correct

In **App2**, (a) is cognitive trust from the human or the perceived trustworthiness of the AI system, (b) is classification of trustworthiness based on Table 1 and (c) is human selection that could be based on observable behaviour, rationality or simply delegation of the responsibility. For the classification of actual trustworthiness, we have divided the human expertise based on the traditional investigation notes (*refer Section 3.2*) as High or Low and AI expertise based on the available location data as High or Low. Furthermore, we have categorized the cases where AI's suggestion and human's diary notes correspond to the correct selection of the hotspot for predictive policing. The hotspot's correctness can never be proven in real life as it would require complete information about all the areas. Hence, to simplify, in our work the correctness of a hotspot was simply based on the permutations of trustworthiness and combinations of AI & Human correctness, refer to Table 1. As to **Calib1** and **Calib2**, the main difference between these two measures of calibrated trust was the agreement percentage. **Calib2** would count the trials in which the participants and the AI's predictions agreed

[7] As opposed to Yang et al [90], we will treat this measure to appropriate reliance as this measure only focuses on engaging or disengaging with the AI system.

[8] Here, actual trustworthiness is an inherent characteristic of a system that subsumes true capabilities of the system in question [78].

and counted as the final decision. In contrast, the switch percentage **Calib1** would only consider cases where they disagreed and had to switch intentionally.

Our experiment aimed for human-in-the-loop collaboration, where participants made decisions assisted by an AI assistant. Participants, after providing informed consent and answering descriptive and exploratory questions from section 4.2.3, were introduced to the AI assistant. The AI assistant was hypothetically trained on the last ten years' crime data in the Netherlands. Finally, they were assigned to one of the four explanation types as per section 4.2.1.

Step 1: *Trial Study* - Participants were tasked with choosing a hotspot for resource allocation, specifically for police patrol. They read investigation notes for guidance and made their selection. The AI assistant then chose a hotspot based on hypothetical training data from https://data.politie.nl/, offering reasoning using one of four explanation configurations. Participants were then prompted to make a final selection, either affirming their or the AI assistant's choice or selecting a different hotspot and providing a reason.

Step 2: *Main Study* - After completing the trial round, the participants received details about the main study, which consists of eight rounds. The participants had a limited 3-minute window for hotspot selection informed by our preliminary study with expert users. In addition, at this step, participants were asked to tick a checkbox if they believed that the AI assistant was better at the task than they themselves. They were instructed that the (*hypothetical*) intelligence unit chief would review their hotspot selections at the end of the experiment. Correct selections earned +10 points, while incorrect ones incurred a -10 point deduction. Due to time constraints, immediate result verification was not possible, prompting participants to proceed to the next round swiftly. Additionally, a bonus was promised for those achieving a top 3 score. Figure 2 describes all the steps performed by the participants.

Step 3: *End of the Study* - Participants completed the post-experiment questionnaire after the main study. They were asked to rate task stakes perception and how familiar they were with the geographical areas shown in the study. Furthermore, we also asked them two open-ended questions: (a) Do you think the AI assistant offered an appropriate explanation of the decision-making process? Why? If not, what explanation do you think it should have offered? and (b) How was your overall experience with this user study? Once they answered these questions, they were shown their final score. A pilot study with five HCI researchers revealed no significant usability flaws, see footnote 5 for details.

4.3 Results

4.3.1 Descriptive Statistics. Of the 12 participants in our user study, 10 had at least a Bachelor's degree. All of them claimed to have know about predictive policing systems, whereas only two of them had heard of or had experience using AI in predictive policing. Our participants' average score of AI literacy was 6.8 (SD = ±1.1) on a scale of 0 to 10, and their propensity to trust predictive policing system was 7.2 (SD = ±1.25) on a scale of 0 to 10. This apriori trust in the system highlights that interface design issues are important to consider as participants in general considered the system legitimate and trustworthy.

Our analysis revealed no major differences in the frequency count of our measures of appropriate trust between the explanation types. However, subjective trust scores were comparatively higher for hybrid explanations (Mean = 60.98, SD = 5.62) when compared to all other explanation types (Mean = 41.23, SD = 10.83).

4.3.2 Qualitative Analysis. We first translated the transcripts in English with the help of two native bilingual speakers and performed qualitative analysis using a reflexive thematic analysis [9]. We inductively generated individual codes from our participants' responses to the open-ended questions and then clustered them into code groups. We identified two main areas: one related to explanation presentation & clarity, and the other related to perceptions of the AI system.

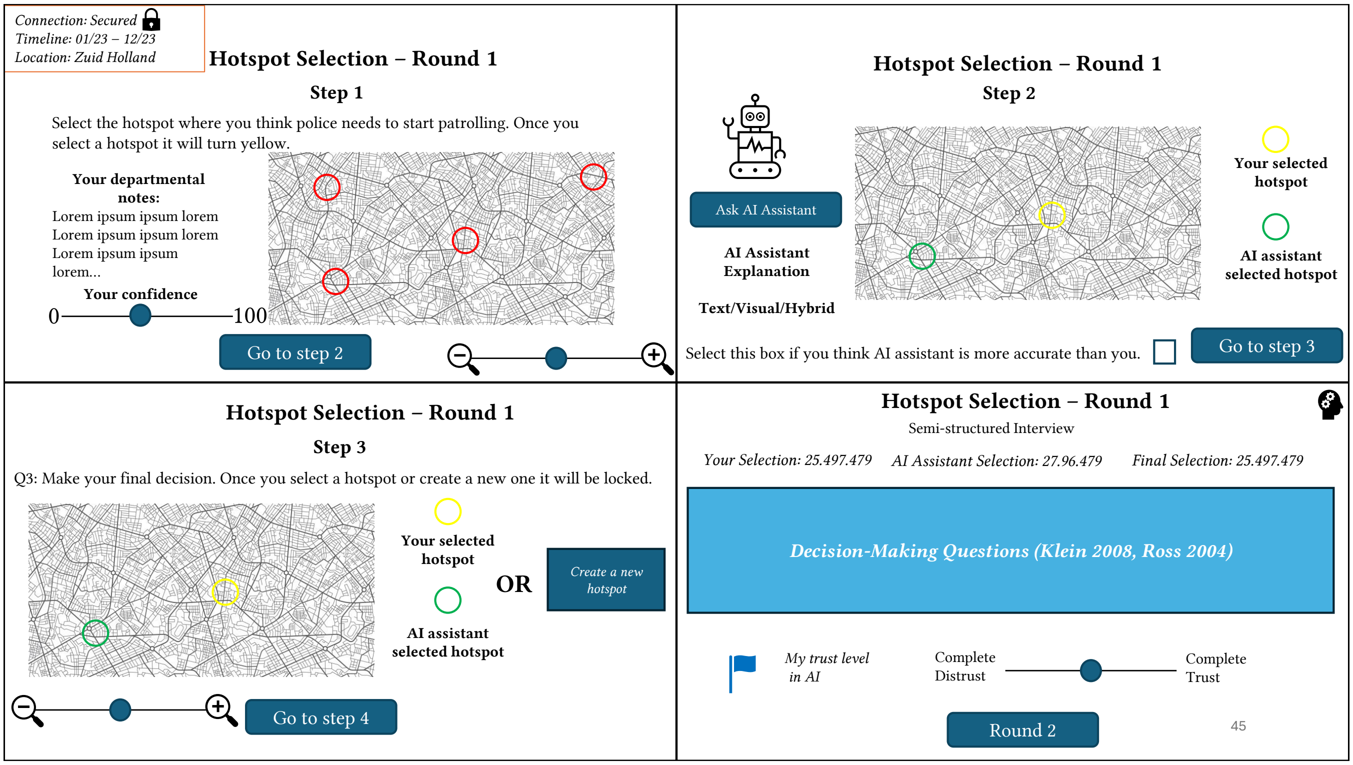
Explanations Presentation and Clarity: Overall, participants from the hybrid and textual explanations group found the explanations to be clear and structured. P7 (textual explanation) wrote, "*the use of public language rather than technical jargon helped decide to go with the AI assistant*". On the other hand, there were mixed reviews from participants for the visual explanations and no explanations categories. P11 (no explanations) expressed the desire for more underpinning or context with the explanation. P5 (visual explanation) found visual explanations to be overwhelming. Also, 50% of participants wrote they followed their reasoning first and then looked at AI assistant's recommendation.

Perceptions of the underlying AI system: There were mixed reviews regarding the help provided by the AI assistant. From the far opposing end, there were concerns about the use of AI in predictive policing where P8 (baseline) wrote, "*the use of AI in predictive policing is fundamentally wrong because you cannot train a system to do policing*". Interestingly, we also found some quotes related to AI capabilities that supported P8's thinking such as "*even explanations will not help in trusting fundamentally biased system*". Also, P8's apriori trust in system was rated at 2 on a scale of 0 to 10, which also supports P8's views. Of the two other participants, who had a rating of 5 or less wrote, "*AI does not possess human intuition and experience. Hence it cannot help in the way that my notes from my teams can*" - P12 (textual) and "*AI rarely captures the considerations of the perpetrator, which is important in understanding any crime as found discussed among officers*" - P4 (Hybrid). Some participants appreciated the AI assistant's decision, when it aligned with their understanding (P1, P7, P10, P15) or provided additional information that was new to them. For example, P1 wrote, "*The information about the existing military unit was useful because it requires a cooperative operation then.*"

5 Second User Study - "Lay Users"

We conducted another user study to understand user expertise's role and assess explanations' role in fostering appropriate trust at a scale. We computed the required sample size using G*Power [25] for an ANOVA with main effects and interactions, specifying the default effect size of 0.25, a significance threshold of α = 0.05, a desired power of 0.8, four groups, and the respective degrees of freedom. The result indicated that we require approximately 179 participants.

We recruited 209 participants from *Prolifc* with an approval rate greater than 95%. Each participant was at least 18 years old,


Connection: Secured
Timeline: 01/23 – 12/23
Location: Zuid Holland

Hotspot Selection – Round 1

Step 1

Select the hotspot where you think police needs to start patrolling. Once you select a hotspot it will turn yellow.

Your departmental notes:
Lorem ipsum ipsum lorem
Lorem ipsum ipsum lorem
Lorem ipsum ipsum
lorem...

Your confidence

Hotspot Selection – Round 1

Step 2

Select this box if you think AI assistant is more accurate than you. □

Hotspot Selection – Round 1

Step 3

Q3: Make your final decision. Once you select a hotspot or create a new one it will be locked.

Hotspot Selection – Round 1

Semi-structured Interview

Your Selection: 25.497.479 AI Assistant Selection: 27.96.479 Final Selection: 25.497.479

Decision-Making Questions (Klein 2008, Ross 2004)

Figure 2: An illustration of the four steps performed by participants of the user-study. In step 1, participants rate their confidence in accurately identifying the hotspot. In step 2, the AI assistant selects a hotspot with its reasoning in form of explanations. In step 3, the participants makes their final decision (Q3). Finally, in step 4, participants rate their subjective trust.

highly proficient in English, and could participate in our study only once. Participants were rewarded based on a $10 hourly rate, and the median completion time was 28 minutes and 11 seconds. Participants were excluded from data analysis if they did not pass at least one of the attention checks. This led to 180 participants (age between 18 and 65+, 94M:86F), *i.e.,* 45 participants per explanation type. The study was conducted on *Qualtrics* in English and was approved by our university’s review Board.

5.1 Methodology

Each participant had to follow the same methodology as the first user study except that they also answered the following question - Have you worked for the police in the past, or are you currently working? Concretely, this means we wanted to filter the participants who have worked for the police in the past or the present as they might have expert knowledge about predictive policing and would not classify as lay users.

5.2 Results

5.2.1 Descriptive Statistics. Of the 180 participants in our user study, 29.44% were between 18 and 24 years old, 44.44% between 25 and 34 years old, 17.77% between 35 and 44 years old, and 8.33% were between 45-65+. 77% of the participants had at least a Bachelor’s degree. None of them claimed neither to have ever worked for the police nor were they aware of the predictive policing system. The average score of AI literacy among participants was 4.67 (SD = ±1.25), and their propensity to trust AI systems was 4.31 (SD = ±0.91). The average duration of the study was 23 minutes (SD = ±4.25), and each participant spent an average of 2 minutes 5 seconds (SD = ±1.03) per round to make the final selection.

5.2.2 Inferential Statistics. Before conducting any statistical analyses, we mapped all (seven-point) Likert scale answers onto an ordinal scale ranging from - 3 (i.e., strongly disagree) to 3 (i.e., strongly agree). The result of Shapiro-Wilk shows that our data followed the normal distribution. Therefore, we conducted an ANOVA with explanations as between-subjects factors and different measures of appropriate trust as the dependent variables. Next to the F statistic and p-value, we also report the partial eta squared η_p^2 effect size. We found no main effect of different explanation types on any measure of appropriate trust ($p > 0.05$, $\eta_p^2 < 0.01$, cohen’s $d = 0.65$).

We conducted another ANOVA with the same between-subjects factors but with subjective trust ratings and usefulness of explanations as the dependent variable. We found a significant difference between different explanation types on the perceived usefulness of the explanations (F (3,1436) = 4.35, $p < 0.005$, $\eta_p^2 = 0.2$, $d = 0.71$). The post hoc analysis revealed that hybrid ($p < 0.013$) and visual explanations ($p < 0.001$) were significantly better than no explanations for the usefulness of explanations ratings. However, we did not find any evidence indicating the effect of different explanation types on subjective trust responses ($p = 0.479$, $\eta_p^2 < 0.01$).

In addition to the analyses described above, we conducted multiple linear regression to analyze the association of independent and dependent variables and exploratory analyses to explore any trends in the data. Our results show that **App2** ($\beta = 4.31$, $p < 0.001$), **Calib2** ($\beta = 8.59$, $p < 0.001$) and AI assistant's correctness ($\beta = 11.01$, $p < 0.001$) predicted the measure **App1** ($R^2 = 0.40$, AIC = 1502, BIC = 1560), with AI assistant's correctness being the strongest predictor. Similarly, we found perceived usefulness of explanations ($\beta = 2.16$, $p < 0.001$) and AI assistant trustworthiness ($\beta = 12.58$, $p < 0.001$) predictors of **App2** ($R^2 = 0.393$, AIC = 1311, BIC = 1314) other than **App1**. We also found **Calib2** ($\beta = 9.34$, $p < 0.001$) and perceived usefulness of explanations ($\beta = 19.45$, $p < 0.001$) predicted the subjective trust scores ($R^2 = 0.267$). Finally, we did not find evidence of any exploratory variable affecting measures of appropriate trust.

5.2.3 Qualitative Analysis. We followed a similar approach as in section 4.3.2 to perform our qualitative analysis. We identified two main topics of interest:

Evaluation of AI's reasoning: Participants, in general, had a positive attitude towards the AI assistant across all explanation types due to (a) lack of expertise for the task, P24 (no explanation): *"I think this system know what it is going, I just need to use it accordingly as this task is very new to me"*, (b) in-depth reasoning of the decision, P96 (textual explanation): *"I believe various factors considered by the AI, such as historical crime data, weather, demographics, and spatial relationships are useful to decide."*, and (c) breaking the tunnel vision, P77 (visual explanation): *"I find visual information appealing and photos, maps, past crime patterns are right to the point, especially the link with weather is something I could never think off."* Some participants expressed reassurance from AI's logical reasoning (P9, P23, P55, P149, P180) and expressed higher trust when their hotspot selection was similar to AI (P112, P106, P155, P47, P33).

Doubts about AI's effectiveness: Several participants (23%) expressed scepticism about the AI assistant's effectiveness irrespective of explanation types. They put forward the desire for consideration of (a) real-time factors (P25, P40, P164), (b) more transparency (P54, P109, P1172), (c) resolution of discrepancies between AI and personal judgement (P37, P111, P140), and (d) providing validation approaches for AI decision-making (P50, P74, P101). Furthermore, 12 participants reported that if the explanation was hard to understand and follow, they just followed the AI assistant's answer because it is too much work to determine whether the AI is right or wrong. E.g., P78: "*region around Assen is under control of military so how as a police officer can make any judgement, go with AI!*"

6 Discussion

6.1 Effect of explanations on appropriate and subjective trust (RQ1 and RQ2)

Our findings show that explanation types, including 'no explanations' had no impact on any measures of appropriate trust in either user study (RQ1). This lack of effect can be understood through the lens of the illusion of explanatory depth [14], a cognitive bias where individuals overestimate their understanding of complex systems. In the context of AI explanations, users could have believed that they comprehend the AI's decision-making process more thoroughly than they actually do, leading to a false sense of understanding regardless of the explanation provided.

To further interpret our results, let's revisit the definitions of our measures, **App1** and **App2**. For **App1** to occur, participants must [not] follow [in]correct recommendations *i.e.,* their appropriate reliance on the system, and for **App2**, understanding both the trustor and trustee's trustworthiness is crucial. Our analyses indicate that, on average, participants correctly selected the hotspot four times in study 1 and three times out of eight rounds in study 2, suggesting a 50% error rate. This high error rate could be attributed to explainability pitfalls, where explanations fail to effectively convey the AI system's reasoning or limitations [17].

Moreover, in user study 1, participants utilized the AI assistant to confirm their intuitive professional judgment rather than comparing trustworthiness, leading to a lack of substantial variations in explanatory formats. Therefore, regardless of expertise, participants failed to perceive meaningful distinctions in trustworthiness, leading us to conclude that there was no effect on appropriate trust, regardless of the type of explanation provided.

For RQ2, our findings indicated that hybrid explanations were rated better on subjective trust than all other explanation types in study 1. However, this trend was not apparent in study 2. This result suggests potential variations in how different user groups perceive and respond to explanation types echoing the work by Szymanski et al. [84]. These differences can be understood through the lens of the ironies of automation, which posit that as systems become more automated, the role of human operators becomes more critical yet potentially more challenging [23].

6.1.1 Increase in trust doesn't mean trust is appropriate. The relationship between increased trust and its appropriateness in AI systems presents a complex challenge that warrants careful examination. Our findings reveal that while hybrid explanations significantly enhanced subjective trust in study 1, this increase did not correlate with improved appropriateness. This disconnect challenges the prevalent assumption that higher trust inherently leads to better outcomes [5], particularly in sensitive domains like predictive policing where the effectiveness of decision-making is paramount.

Appropriate trust represents a multifaceted construct influenced by context, agent characteristics, and underlying cognitive processes [13]. Our research suggests that while explanations may boost user trust in AI systems, this enhanced trust does not necessarily translate to improved decision-making capabilities. This finding presents two potential paths forward: exploring alternative trust-building approaches or addressing potential deficiencies in explanation quality [5].

Current research presents divergent perspectives on addressing this challenge. Some researchers advocate for diversifying trust-building methods through transparency, user engagement, and iterative feedback [27, 86], while others emphasize the need to enhance explanation quality and clarity [57]. Based on our findings, we recommend a balanced approach that evaluates trust through multiple lenses: appropriateness (goal alignment), system purpose (usability), and user requirements (usefulness). Additionally, we suggest exploring Miller's Evaluative AI framework as a potential alternative to traditional XAI approaches, offering hypothesis-driven decision support [65].

6.2 Usefulness of explanations, role of user expertise and exploratory measures

Our analysis reveals several key findings regarding the relationship between explanations, trust, and user expertise in AI systems. Perceived usefulness of explanations significantly influenced subjective trust scores (RQ3), though this did not translate to appropriate trust formation. The correlation between explanation usefulness and subjective trust suggests that users' trust assessments incorporate both prediction accuracy and the perceived value of explanations.

While user expertise did not moderate the role of explanations in building appropriate trust (RQ4), expert users showed significantly higher subjective trust with hybrid explanations. This aligns with existing research emphasizing the importance of user expertise in AI system trust [66, 74, 84] and supports previous arguments for expertise-tailored explanations [4, 12, 30, 83]. In addition, visual and hybrid conditions were linked to perceived usefulness in study 2, but the results were opposite in study 1, again possibly hinting at the role of user expertise. A marked distinction emerged in decision-making patterns between user groups. Lay users predominantly followed AI recommendations directly (42%), while expert users demonstrated a more analytical approach (75%) through their open-text responses. This difference aligns with Wang et al.'s findings [88] regarding inexperienced users' susceptibility to reinforcement effects, particularly evident when lay users encountered unfamiliar information.

Our exploratory variables analysis revealed positive correlations between AI trust propensity, AI literacy, subjective trust scores, and crime classification in study 1. However, these correlations did not persist in study 2, suggesting the influence of contextual factors. This variance might be attributed to differences in participant characteristics (study 1 showed higher average AI literacy scores of 6.80 compared to 4.67 in study 2) and system-specific factors in predictive policing. The identified predictors of appropriate trust, particularly AI assistant correctness and trustworthiness, demonstrated consistency across both expert and lay user contexts.

6.2.1 *Policy measures for public sector AI systems.*

Public sector AI systems, particularly in predictive policing, face critical temporal and spatial challenges that require careful consideration before development or deployment. Our research identifies three key challenges: confirmation bias in AI interpretation across both expert and non-expert users, alignment with pre-existing biased judgments which risks amplifying systemic biases [80], and the concerning disconnect between increased subjective trust and actual decision quality when explanations are provided.

These findings carry significant policy implications for responsible AI implementation. Most notably, the observation that increased subjective trust through explanations does not correlate with improved decision-making necessitates a strategic shift in AI development approaches. We propose three essential policy measures that extend to broader XAI systems: implementing performance metrics focused on decision quality rather than subjective trust; incorporating user expertise levels into system design, given the observed variations between user groups; and preserving human discretion in decision-making, as evidenced by expert users' successful integration of AI advice with professional knowledge. This last point is particularly relevant when compared to systems like Germany's, where deviation from AI recommendations is constrained [62].

6.3 Limitations and Future Work

Our study faced several notable limitations that should inform the interpretation of results and future research directions. The primary constraint was Study 1's small sample size, which, while consistent with sampling methods used in comparable predictive policing research [28], potentially limits the generalizability of our findings to broader populations. The study's experimental design, utilizing hypothetical scenarios focused on individual hotspot selection with AI assistance, represents a simplified version of actual police decision-making processes, which typically involve team-based approaches. This aligns with Ferguson's observations [26] regarding "big data policing," where concerns often center more on underlying data quality than system trust. The use of experimental conditions and fictitious vignettes, while methodologically necessary, couldn't fully capture the complexity of authentic police decision-making environments. To address these limitations, future research should employ larger sample sizes and explore alternative XAI approaches to better understand appropriate trust development. More robust methodological approaches could include virtual reality simulations of police decision-making [55], deliberative polling techniques, or real-world interventions. These enhanced approaches would provide more comprehensive insights into the complexities of trust formation in predictive policing systems.

7 Conclusion

In this paper, we looked at the effect of different types of explanations (text, visual, and hybrid) and user expertise (retired police officers and lay users) on fostering appropriate trust in an AI-based predictive policing system. Our results show that a hybrid form of explanations raised the subjective trust of expert users compared to lay users in the AI system. However, none of the explanation types helped participants in forming appropriate trust in the system. We argue that this result of an increase in trust is worrisome, as it does not lead to better decisions. Based on these results, we highlight challenges in building appropriate trust in human-AI interaction and propose important policy recommendations centered around fostering appropriate trust in AI-based predictive policing systems. We hope this paper will serve as a "call to action" for the UMAP community to shift focus from the use of explanations for just promoting trust in AI systems to fostering appropriate trust instead.

Acknowledgments

This research was (partly) funded by the Hybrid Intelligence Center, a 10-year programme funded the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientiic Research, grant number 024.004.022, by TAILOR Connectivity Fund Grant through EU H2020 program under agreement number 952215 and by EU H2020 ICT48 project "Humane AI Net" under contract 952026.

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 838–849.

[3] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. 2022. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law* (2022), 1–17.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–16.

[6] Gabriel Diniz Junqueira Barbosa, Dalai dos Santos Ribeiro, Marisa do Carmo Silva, Hélio Lopes, and Simone Diniz Junqueira Barbosa. 2022. Investigating the relationships between class probabilities and users' appropriate trust in computer vision classifications of ambiguous images. *Journal of Computer Languages* 72 (2022), 101149.

[7] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2022. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 32, 1 (2022), 110–138.

[8] Astrid Bertrand, James R Eagan, and Winston Maxwell. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 943–958.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. doi:10.1145/3449287

[11] Justin B Bullock. 2019. Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration* 49, 7 (2019), 751–761.

[12] Margaret Burnett. 2020. Explaining AI: fairly? well?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces.* 1–2.

[13] Jessie YC Chen and Michael J Barnes. 2014. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (2014), 13–29.

[14] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21).* Association for Computing Machinery, New York, NY, USA, 307–317. doi:10.1145/3397481.3450644

[15] EU Commission. 2021. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Proposal for a regulation of the European parliament and of the council.

[16] David Danks. 2019. The value of trustworthy AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 521–522.

[17] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. 2022. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government information quarterly* 39, 2 (2022), 101666.

[18] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

[19] Chadha Degachi, Siddharth Mehrotra, Mireia Yurrita, Evangelos Niforatos, and Myrthe Lotte Tielman. 2024. Practising Appropriate Trust in Human-Centred AI Design. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24).* Association for Computing Machinery, New York, NY, USA, Article 269, 8 pages. doi:10.1145/3613905.3650825

[20] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces.* 275–285.

[21] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[22] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19).* Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312787

[23] Mica R Endsley. 2023. Ironies of artificial intelligence. *Ergonomics* 66, 11 (2023), 1656–1668.

[24] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency.* PMLR, 160–171.

[25] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[26] Andrew Guthrie Ferguson. 2018. The legal risks of big data policing. *Crim. Just.* 33 (2018), 4.

[27] Carolina Ferreira Gomes Centeio Jorge, Siddharth Mehrotra, Myrthe L. Tielman, and Catholijn M. Jonker. 2021. Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams. *Proceedings of the 22nd International Workshop on Trust in Agent Societies, London, UK* (2021).

[28] Monica M Gerber and Jonathan Jackson. 2017. Justifying violence: legitimacy, ideology and public support for police use of force. *Psychology, crime & law* 23, 1 (2017), 79–95.

[29] Dominik Gerstner. 2018. Predictive policing in the context of residential burglary: An empirical illustration on the basis of a pilot project in Baden-Württemberg, Germany. *European Journal for Security Research* 3, 2 (2018), 115–138.

[30] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. 2021. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters* 2, 4 (2021), e61. doi:10.1002/ail2.61

[31] Wim Hardyns and Anneleen Rummens. 2018. Predictive policing as a new tool for law enforcement? Recent developments and challenges. *European journal on criminal policy and research* 24 (2018), 201–218.

[32] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2022. It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 89–101.

[33] Fred Hohman, Arjun Srinivasan, and Steven M Drucker. 2019. TeleGam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS).* IEEE, 151–155.

[34] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 624–635.

[35] Anthony Jameson, Bettina Berendt, Silvia Gabrielli, Federica Cena, Cristina Gena, Fabiana Vernero, Katharina Reinecke, et al. 2014. Choice architecture for human-computer interaction. *Foundations and Trends® in Human–Computer Interaction* 7, 1–2 (2014), 1–235.

[36] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. doi:10.1145/3491102.3517533

[37] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–14.

[38] Mohammad T Khasawneh, Shannon R Bowling, Xiaochun Jiang, Anand K Gramopadhye, and Brian J Melloy. 2003. A model for predicting human trust in automated systems. *Origins* 5 (2003).

[39] Been Kim, Caleb M Chacha, and Julie A Shah. 2015. Inferring team task plans from human meetings: A generative modeling approach with logic-based prior. *Journal of Artificial Intelligence Research* 52 (2015), 361–398.

[40] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV).* 563–578.

[41] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19).* Association for Computing Machinery, New York, NY, USA, 29–38. doi:10.1145/3287560.3287590

[42] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16).* Association for Computing Machinery, New York, NY, USA, 1675–1684. doi:10.1145/2939672.2939874

[43] Retno Larasati, Anna De Liddo, and Enrico Motta. 2023. Meaningful Explanation Effect on User's Trust in an AI Medical System: Designing Explanations for Non-Expert Users. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–39.

[44] Retno Larasati, Anna De Liddo, and Enrico Motta. 2023. Meaningful Explanation Effect on User's Trust in an AI Medical System: Designing Explanations for

Non-Expert Users. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 30 (dec 2023), 39 pages. doi:10.1145/3631614

[45] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[46] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 369 (oct 2023), 22 pages. doi:10.1145/3610218

[47] Matthias Leese. 2024. Staying in control of technology: Predictive policing, democracy, and digital sovereignty. *Democratization* 31, 5 (2024), 963–978.

[48] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. doi:10.1145/3531146.3533182

[49] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[50] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 580, 15 pages. doi:10.1145/3491102.3517606

[51] Vidushi Marda and Shivangi Narayan. 2020. Data in New Delhi's predictive policing system. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 317–324.

[52] Brendan Max. 2022. SoundThinking's Black-Box Gunshot Detection Method: Untested and Unvetted Tech Flourishes in the Criminal Justice System. *Stan. Tech. L. Rev.* 26 (2022), 193.

[53] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[54] Maranda McBride and Shona Morgan. 2010. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions* (2010), 1–11.

[55] John McDaniel and Ken Pease. 2021. *Predictive policing and artificial intelligence*. Routledge.

[56] John M McGuirl and Nadine B Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors* 48, 4 (2006), 656–665.

[57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[58] S. Mehrotra. 2024. *Designing for Appropriate Trust in Human-AI Interaction*. Dissertation (TU Delft). Delft University of Technology. doi:10.4233/uuid:5a0c475b-5494-4f7a-a91c-796975233d95

[59] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM J. Responsib. Comput.* 1, 4, Article 26 (Nov. 2024), 45 pages. doi:10.1145/3696449

[60] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. 2023. Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations. In *Volume 368: HHAI 2023: Augmenting Human Intellect*. 436–439. doi:10.3233/FAIA230121

[61] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. Integrity-based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Trans. Interact. Intell. Syst.* 14, 1, Article 4 (jan 2024), 36 pages. doi:10.1145/3610578

[62] Albert Meijer, Lukas Lorenz, and Martijn Wessels. 2021. Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review* 81, 5 (2021), 837–846.

[63] Albert Meijer and Martijn Wessels. 2019. Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration* 42, 12 (2019), 1031–1039.

[64] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[65] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 333–342.

[66] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *ArXiv* abs/1811.11839 (2018). https://api.semanticscholar.org/CorpusID:54087635

[67] Ishmael Mugari and Emeka E Obioha. 2021. Predictive policing and crime control in the United States of America and Europe: Trends in a decade of research and the future of predictive policing. *Social sciences* 10, 6 (2021), 234.

[68] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27 (2017), 393–444.

[69] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLoS ONE* 15, 2 (2020).

[70] Serena Oosterloo, Gerwin van Schie, Jo Bates, Paul Clough, Robert Jäschke, Jahna Otterbacher, et al. 2018. The politics and biases of the "crime anticipation system" of the Dutch police. In *Proceedings of the international workshop on bias in information, algorithms, and systems*, Vol. 2103. CEUR WS, 30–41.

[71] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G Jentsch. 2013. Building appropriate trust in human-robot teams. In *2013 AAAI spring symposium series*.

[72] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[73] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8779–8788.

[74] Mireia Ribera and Àgata Lapedriza García. 2019. Can we do better explanations? A proposal of user-centered explainable AI. CEUR Workshop Proceedings.

[75] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233.

[76] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–108.

[77] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[78] Nadine Schlicker and Markus Langer. 2021. Towards Warranted Trust: A Model on the Relation Between Actual and Perceived System Trustworthiness. In *Proceedings of Mensch Und Computer 2021* (Ingolstadt, Germany) *(MuC '21)*. Association for Computing Machinery, New York, NY, USA, 325–329. doi:10.1145/3473856.3474018

[79] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1616–1628.

[80] Friso Selten, Marcel Robeer, and Stephan Grimmelikhuijsen. 2023. 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review* 83, 2 (2023), 263–278.

[81] R Jay Shively, Joel Lachter, Summer L Brandt, Michael Matessa, Vernol Battiste, and Walter W Johnson. 2018. Why human-autonomy teaming?. In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2017 International Conference on Neuroergonomics and Cognitive Engineering, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*. Springer, 3–11.

[82] Auste Simkute, Ewa Luger, Mike Evans, and Rhianne Jones. 2020. Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS' 20 Companion)*. Association for Computing Machinery, New York, NY, USA, 263–268. doi:10.1145/3393914.3395862

[83] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Duen Horng Chau, Alex Endert, and Daniel Keim. 2020. Should we trust (x) AI? Design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433* (2020).

[84] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[85] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[86] Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra, and Myrthe Tielman. 2023. Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. *European Journal of Work and Organizational Psychology* (2023), 1–14.

[87] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. doi:10.1145/3579605

[88] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[89] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[90] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?.

In *Proceedings of the 25th international conference on intelligent user interfaces.* 189–201.

[91] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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.* 295–305.