Human Response to Decision Support in Face Matching: The Influence of Task Difficulty and Machine Accuracy

Marina ESTÉVEZ-ALMENZAR [a,1], Ricardo BAEZA-YATES [a,b] and Carlos CASTILLO [a,c]

[a] *Universitat Pompeu Fabra*
[b] *ICREA*
[c] *Northeastern University*

ORCiD ID: Marina Estévez-Almenzar https://orcid.org/0009-0005-8813-8593, Ricardo Baeza-Yates https://orcid.org/0000-0003-3208-9778, Carlos Castillo https://orcid.org/0000-0003-4544-0416

Abstract. Decision support systems enhanced by Artificial Intelligence (AI) are increasingly being used in high-stakes scenarios where errors or biased outcomes can have significant consequences. In this work, we explore the conditions under which AI-based decision support systems affect the decision accuracy of humans involved in face matching tasks. Previous work suggests that this largely depends on various factors, such as the specific nature of the task and how users perceive the quality of the decision support, among others. Hence, we conduct extensive experiments to examine how both task difficulty and the precision of the system influence human outcomes. Our results show a strong influence of task difficulty, which not only makes humans less precise but also less capable of determining whether the decision support system is yielding accurate suggestions or not. This has implications for the design of decision support systems, and calls for a careful examination of the context in which they are deployed and on how they are perceived by users.

Keywords. Decision support systems, face matching, human factors.

1. Introduction

Decision support systems are a key modality of use of Artificial Intelligence (AI). They can be categorized by the extent to which the final decisions depend on them, from having no influence whatsoever, to being fully autonomous, with most cases operating somewhere between these two extremes [1,2]. Ideally, these systems together with human operator(s) create a hybrid human-machine intelligence that exploits the expertise of human operators with the capacity to find patterns in historical data that yield better decisions. This is particularly critical in applications that have significant effects on people, such as those described as *high risk* by the European AI Act (EU Regulation 2024/1689) [3].

[1] Corresponding Author: Marina Estévez-Almenzar, e-mail: marina.estevez@upf.edu.

In these applications, given the ethical and legal requirements for human oversight, it is unlikely that fully automated systems are deployed in the near future. Instead, hybrid systems combining human and machine intelligence are likely to become the norm.

When humans and machines work together, they should be evaluated together [4]. However, human-algorithmic behavior involves complex emerging patterns, uncertainties, and a certain degree of unpredictability that is in tension with the goal of developing safe and trustworthy systems. Modeling how human operators respond to recommendations produced by an algorithm is paramount.

This is an active research topic, and previous work (surveyed in §2) investigates aspects such as the nature of the task for which decision support is provided, the way in which machine assistance is framed, preconceptions that make users averse or overreliant on algorithms, and the accuracy or perceived accuracy of the algorithmic suggestions, among other factors.

In this work, we study a face matching scenario in which a person is tasked with determining whether two photos correspond to the same person, and uses a face matching system as a decision support tool. Our work addresses a series of research questions (§3) related to the extent to which a decision support system can enhance human accuracy, focusing on responses to both correct and incorrect suggestions and the effects of fluctuating (increasing or decreasing) system accuracy.

We tackle these questions through a series of experiments (§4) conducted via crowdsourcing in which experimental variables include task difficulty, decision support accuracy, and whether a notification is given to users when decision support accuracy might change.

Our results and discussion (§5-§6) show a strong influence of task difficulty, which not only makes human annotators less accurate, but also makes them more prone to be misled by inaccurate decision support and less aware of the accuracy of different decision support systems. We also find important differences on human perceptions when decision support errors appear at the beginning or end of the sequence of tasks, compared to when they are randomly distributed throughout the sequence of tasks. The last section (§7) presents our conclusions and outlines future work.

2. Related Work

2.1. Decision Support: The Nature of the Task

Previous work highlights how the nature of the task plays a crucial role in human-machine interaction with a decision support system. One of the main axes of analysis is the distinction between objective and subjective tasks. It has been shown that the more objective the task is perceived to be, the more likely the human is to be influenced by the machine [5]. However, we also find evidence that in some cases the distinction between objective and subjective tasks does not play an important role in how people are influenced by machine suggestions [6]. This apparent contradiction seems to be reconciled in the work of Hou *et al.* [7], where the authors suggest that the really influential factor is the machine competence perceived by the human. This, in turn, depends heavily on how the decision support agent is presented: different framings of the same agent can shape different human perceptions, leading to inconsistent outcomes. In a similar vein,

Mahmud *et al.* [8] highlight the moral nature and the complexity of the task at hand. People tend to move away from the machine when it comes to making moral decisions, such as those related to legal or medical issues [9,10,11]. Also, people tend to reject the machine when it comes to tasks that do not require high computational skills [12,13]. Furthermore, it has been noted that when utilitarian results hold significant value, there is a preference for AI recommenders instead of human ones, whereas when hedonic aspects are prioritized, there tends to be a resistance to AI recommenders in favor of human decisions [14].

2.2. Behavioral Patterns in Human-Machine Interaction

Several patterns of behavior that emerge from the interaction between a human and a machine have been extensively studied. *Algorithmic aversion* is defined as a negatively biased perception of algorithms that manifests itself in a behavior of rejection towards the algorithm with respect to human agents. This aversion is especially reinforced when the human interacting with the machine observes that the machine makes mistakes [15]. In contrast, in the evaluation of hybrid resolutions of moral problems humans tend to be evaluated more leniently than the machine, which is known as a human self-interest bias [16]. Conversely, *algorithmic appreciation* is known as a positive dominance of the algorithm that helps users avoid mistakes. However, this dominance might be detrimental if humans become overly dependent on machine outcomes, which may lead to errors [17]. Another cognitive bias, well-recognized in psychology but less explored within the realm of human-machine interaction, is the tendency of humans to prefer information that aligns with their existing beliefs [18].

2.3. Human-Machine Interaction in Facial Recognition

Prior decisions made by face recognition systems influenced subsequent face matching decisions made by human operators. When face pairs were incorrectly labeled by the machine, the precision of humans decreased by drawing attention away from face images, even when humans were warned that machine predictions could be inaccurate [19]. Furthermore, decision-deferral rates in human-machine systems influence both human performance and trust during face-matching tasks [20]. It is well documented that human face recognition accuracy can be improved by the wisdom of crowds: combined judgment of many is better than the decision of an individual [21]. There is a similar benefit to merging the performance of multiple algorithms [22]. When considering decisions resulting from the fusion of human decision and machine decision, the results can lead to large performance improvements compared to the human response or the algorithm response alone [23].

Our work addresses task complexity, a subject that that has not been explored in as much detail as other aspects in the surveyed literature [24], positioning it as a pivotal element to be carefully considered in the design of decision support systems. Additionally, in many real-world scenarios in which data evolves, and given that machine learning models should be trained and applied on data with identical distributions, keeping models up to date is a critical task [25,26,27]. Updating the models produces variability in performance, causing an effect on interaction patterns [28]. As far as we know, previous research has not thoroughly examined the effects of potential machine variability

on human-machine interactions, nor has it been thoroughly studied the effects of how errors are distributed along a sequence of tasks in machines of equal average accuracy. In this paper we consider machines of varying accuracy and observe how these variations affect human performance. We also study whether notifying the human operator each time a variation in the machine occurs makes any difference in joint human-machine performance.

3. Research Questions

Our experiments are designed to address the following research questions:

RQ1 *Does an AI-based decision support system improve human performance in a face matching task?*
We test to what extent the support of a high-accuracy machine improves human accuracy in solving a face-matching task, while testing whether this improvement depends on the difficulty of the task.

RQ2 *Does a low accuracy AI decision support system improve or deteriorate human performance in a face matching task?*
We test to what extent the support of a low-accuracy machine deteriorates human accuracy, while testing whether this deterioration depends on the difficulty of the task.

RQ3 *Does a variable accuracy AI decision support system improve or deteriorate human performance in a face matching task?*
We test to what extent the support of a variable-accuracy machine improves or deteriorates human accuracy, while testing whether this change depends on the difficulty of the task. We also test whether this change depends on human awareness of this machine variability.

4. Experimental Setup

To investigate our research questions, we designed three experiments. We considered three independent variables: (i) problem difficulty, (ii) machine accuracy, and (iii) change notifications. *Problem difficulty*, described in §4.1, indicates the difficulty of the tasks assigned to the participant. *Machine accuracy*, described in §4.2, indicates the accuracy of the decision support assigned to the participant. *Change notification* indicates whether the participant is notified or not when the decision support system changes, and only applies to the experiment in which the accuracy varies. We measured three dependent variables: accuracy, influence factor, and confirmation factor. The three dependent variables are defined in §4.3.

Experiment 1: With / Without Decision Support This experiment works as our "control experiment" and is designed for the purpose of investigating RQ1. Two groups of different participants were compared. The control group solved the task without machine suggestions. The experimental group of participants received, for every task, a suggestion from a system having 95% accuracy.

Experiment 2: With Degraded Decision Support This experiment is designed for the purpose of investigating RQ2. Two experimental groups were compared. One of them solved the face matching tasks while receiving suggestions from a 5% accuracy machine and the other one while receiving suggestions from a 50% accuracy machine.

Experiment 3: With Variable Decision Support Machine This experiment is designed for the purpose of investigating RQ3. Four groups of different participants were compared. Two groups solved the tasks while receiving suggestions from an *increasing* accuracy machine (that we note INC machine). One of these groups was notified every time the machine changed, and the other group was not. Conversely, two other groups solved the tasks while receiving suggestions from a *decreasing* accuracy machine (DEC), and similarly, one of them was notified of changes, while the other was not.

Our research plan was reviewed and approved by the Ethics Review Board of the university of the lead author. The review included compliance with internationally accepted ethical principles in research, and with personal data protection, in particular by the EU General Data Protection Regulation (2016/679).

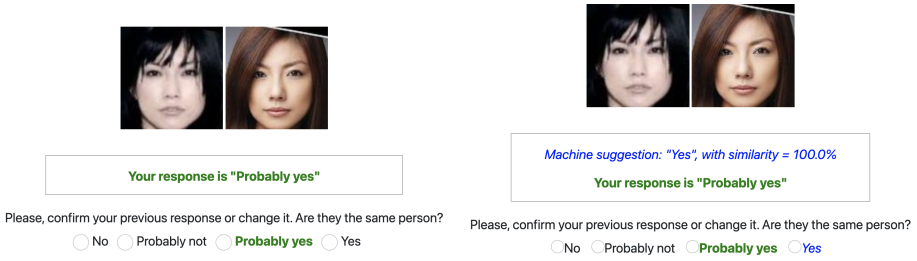
4.1. Procedure

We performed an online user study, with the following structure.

Participant recruitment We recruited participants through a crowdsourcing platform for experimentation named Prolific.[2] We considered four countries in continental Europe in which Prolific has large user bases: France, Germany, Italy, and Spain, plus the United Kingdom. We made sure that our sets of participants were gender balanced, a feature that the platform provides based on participants' disclosures of gender. In total, we recruited 320 participants, and each participant annotated 30 different pairs of images. In total, we collected 9,600 participant's annotations, and a total of 60 different pairs were annotated under various conditions. Participants were paid 9.16 EUR per hour. Specifically, they were paid 1,07 EUR for labeling 30 pairs of images, with an average completion time of 10 minutes. To encourage participants' effort, we set a bonus: participants were warned (and reminded during the study) that if they managed to correctly match more than 80% of the pairs (more than 24 pairs) they would receive an extra payment of 30%.

Tasks selection Each of the experiments was carried out twice. Once with a set of pairs that we call *Normal Set*, and once with a different set of pairs that we call *Hard Set*. Both sets consist of pairs from the DemogPairs testing database, and their classification as *Normal* or *Hard* is based on a preliminary study in which a total of 540 pairs were tested by a total of 162 participants (10 different pairs per participant, 3 different participants per pair). These participants were paid the same as described previously. In this study, participants rated these pairs by answering the question 'Are they the same person?' with the options "No", "Probably not", "Not sure", "Probably yes", or "Yes". This allows us to categorize some of these pairs according to the difficulty experienced by participants in solving the task. In this previous study, these 540 pairs were also evaluated by two state-of-the-art models jointly. These two models are IR50+ArcFace and LightCNN, and the joint accuracy over DemogPairs test dataset is above 95%. By "joint accuracy" we mean the accuracy based on the mean response of both models. This ensemble machine is the base decision support system we use in our experiments.

[2] www.prolific.co



(a) Control group (b) Experimental groups

Figure 1. Survey screenshots.

- Hard Set: We selected those pairs whose mean participant response was very close to "Not sure", and at least one of the three participants made a mistake. We also selected those pairs where the mean response is exactly "Not sure". We obtained a total of 69 pairs (63 positive and 6 negative). After a careful manual inspection, from among the 63 positive pairs we chose the 24 most difficult ones which together with the 6 negative ones (with no occlusion *i.e.*, no objects obstructing parts of the face) form the *Hard Set*.
- Normal Set: We define another set of pairs with a slightly more relaxed human certainty condition: we selected those pairs whose mean response is close to "Not sure", avoiding those whose mean is exactly "Not sure". From these pairs we randomly choose 30 pairs of images, maintaining the above ratio of 24 positive and 6 negative pairs, and avoiding repetitions with the "Hard Set". This left us with 30 pairs that will form the *Normal Set*.

Face matching tasks Participants evaluated one pair of images at a time. Given a pair p, the participant had to answer the question *Are they the same person?*, with the possible options: *No*, *Probably not*, *Probably yes*, or *Yes*. After answering, if the participant was assigned to one of the experimental groups, the machine suggestion was shown together with the machine similarity score associated with the pair s_p on which the suggestion was based: $0 \leq s_p \leq 0.25$ with suggestion *No*, $0.25 < s_p \leq 0.50$ with suggestion *Probably not*, $0.50 < s_p \leq 0.75$ with suggestion *Probably yes* or $0.75 < s_p \leq 1.00$ with suggestion *Yes*. They had the possibility to modify their answer (see Figure 1b). Participants in the control experiment also had the possibility to modify their answer (see Figure 1a).

Exit survey After evaluating the 30 pairs, the participants who interacted with a machine completed an exit survey. They were asked whether the suggestions of the machine had been useful to 1. *make a decision when they were not sure about their answer*, 2. *make a decision when some of the images were blurry or the quality was not good*, 3. *make decisions faster*, 4. *confirm their decision when they were certain about it*, 5. *make more accurate decisions*, and 6. *make them feel more confident about the answer*. Participants had the possibility to answer *Strongly disagree* / *Disagree* / *Neither agree nor disagree* / *Agree* / *Strongly agree*.

4.2. Machines

To simulate machines that adapt to the circumstances that we want to reproduce in each experiment, we introduce noise into the ensemble machine. Given a pair p, we obtain the associated similarity score s_p, with $0 \leq s_p \leq 1$. We define the noise as $f(s_p) = 1 - s_p$, which gives us a noisy similarity score that forces the opposite response. The number of pairs to which this noise is applied depends on the machine to be simulated. These machines are described below.

Static machines We simulate three machines: 95%, 50%, and 5% accuracy machines. Each of them is a realization of the probabilistic situations we simulate, which means that they have exactly the target accuracy in every experiment.

Variable machines We simulate two variable accuracy machines: 1) *INC machine* simulates a model that increases its accuracy over time, and is defined as the concatenation of three machines: 5% - 50% - 95% accuracy machines, each of them solving one third of the total number of tasks. 2) *DEC Machine* simulates a model that decreases its accuracy over time, and is defined as the concatenation of three machines: 95% - 50% - 5% accuracy machines, each of them solving one third of the total number of tasks.

Regarding change notification, there were two conditions: without notification, and with notification. Participants in the first condition were not told anything about variable machine accuracy. Participants in the notification condition, before starting the survey, were shown the following message: *There are three AIs, named machine A, machine B, and machine C. We will notify you every time there is a change*. The notification message was: *Next, you will receive suggestions from machine* $\{A\ /\ B\ /\ C\}$.

4.3. Measurements

Participant accuracy We compute the fraction of correct responses, with respect to the ground truth, given by the participant before (*initial accuracy*) and after (*final accuracy*) seeing the machine suggestion. In the analysis we show the macro-average across participants of the initial and final accuracy, together with their standard deviation.

Interaction Factors Given a pair p solved by a participant, let r_i be the *participant's initial response*, r_f the *participant's final response*, and m the *machine suggestion*[3].

- Influence Factor: We measured how much the machine suggestion influenced the participant's response. This value, that we note IF, ranges from -1 to 1, and is based on the influence defined by Hou *et al.* in [7]. We define

$$IF(r_i, m, r_f) = \begin{cases} -|r_f - r_i| & \text{if } |m - r_i| < 1 \\ \frac{r_f - r_i}{m - r_i} & \text{if } |m - r_i| \geq 1 \end{cases}$$

  In the analysis we show the macro-average of the influence factor. We also measure the probability that this influence is positive, $P(IF > 0)$, zero $P(IF = 0)$, or negative $P(IF < 0)$. We find this metric easy to interpret as, for example, a positive value is given to cases where the participant changes their response in the direction of what is recommended by the machine.

[3] $r_i, r_f \in \{-2, -1, 1, 2\} \equiv \{$*No, Probably not, Probably yes, Yes*$\}, m = -2 + 4s_p \in [-1, 1]$ with $s_p \in [0, 1]$

Table 1. Average participant initial accuracy a_i, final accuracy a_f, difference among both δ, influence factor IF, probability that this influence is positive $P(IF > 0)$, probability that this influence is neutral $P(IF == 0)$, probability that this influence is negative $P(IF < 0)$, and probability of confirmation $P(C)$ for all the experiments with the *Normal Set* and the *Hard Set*. There are 20 participants for every row. (n) stands for *with notification*, (-) for *with no notification*.

Normal Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
no machine	0.67 ± 0.23	0.67 ± 0.23	0	-	-	-	-	-
95%	0.71 ± 0.20	0.80 ± 0.20	+0.09	0.01	0.13	0.78	0.09	0.57
50%	0.69 ± 0.21	0.66 ± 0.21	-0.03	0.05	0.14	0.81	0.05	0.40
5%	0.64 ± 0.28	0.61 ± 0.30	-0.03	-0.03	0.07	0.87	0.06	0.32
INC (n)	0.64 ± 0.21	0.64 ± 0.22	0	+0.00	0.11	0.82	0.07	0.45
INC (-)	0.64 ± 0.21	0.63 ± 0.18	-0.01	+0.00	0.12	0.79	0.09	0.47
DEC (n)	0.70 ± 0.21	0.69 ± 0.21	-0.01	0.04	0.11	0.85	0.04	0.42
DEC (-)	0.71 ± 0.18	0.67 ± 0.21	-0.04	0.10	0.22	0.70	0.08	0.39
Hard Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
no machine	0.57 ± 0.20	0.57 ± 0.20	0	-	-	-	-	-
95%	0.58 ± 0.23	0.65 ± 0.22	+0.07	0.13	0.15	0.81	0.04	0.34
50%	0.55 ± 0.24	0.56 ± 0.22	+0.01	0.24	0.23	0.75	0.02	0.34
5%	0.54 ± 0.22	0.38 ± 0.24	-0.16	0.21	0.22	0.74	0.04	0.26
INC (n)	0.55 ± 0.15	0.56 ± 0.15	+0.01	0.18	0.18	0.78	0.04	0.35
INC (-)	0.49 ± 0.17	0.49 ± 0.17	0	0.17	0.18	0.80	0.02	0.35
DEC (n)	0.58 ± 0.17	0.59 ± 0.17	+0.01	0.11	0.15	0.80	0.05	0.36
DEC (-)	0.53 ± 0.16	0.52 ± 0.18	-0.01	0.13	0.19	0.72	0.09	0.30

- Confirmation Probability: We measured the probability that the participant's initial response and the machine's suggestion match and the participant does not change their final response. This event, that we note as C, occurs when $r_i = r_f$ and $|m - r_i| \leq 1$. In the analysis we show the probability that this occurs, that we note as $P(C)$.

5. Results

Table 1 summarizes our results, which we explain next.

5.1. Experiment 1: With vs. Without Decision Support Machine

For every set of pairs, we compare two groups of participants: those who did not receive machine suggestions and those who received suggestions from the 95% accuracy machine.

Normal Set Almost all participants with no machine suggestions maintained their initial response when given the opportunity to modify it in the vast majority of tasks, making both the initial and final accuracy 0.67 ± 0.23. Results from participants interacting with the 95% accuracy machine suggest that the support of a high accurate machine improves human performance (see left plot in Figure 2) even when the influence is low.

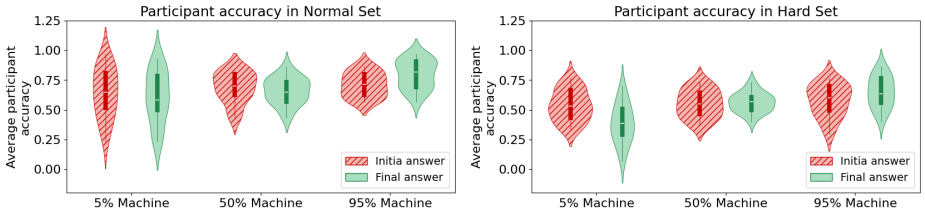


Figure 2. Participant initial and final average accuracy distributions for the *Normal Set* and the *Hard Set*, for 5%, 50%, and 95% Machines in Experiment 1.

Hard Set Similarly, almost all participants with no machine suggestions maintained their initial response when given the opportunity to modify it in the vast majority of tasks, making both the initial and final accuracy 0.57 ± 0.20. As before, participants interacting with the 95% accuracy machine got to improve their accuracy (see right plot in Figure 2). For these participants, the influence is higher than for those in the Normal Set.

Participants perceive the 95% accuracy machine more positively in the Normal Set, as shown in Figure 4, despite its higher influence in the Hard Set.

5.2. Experiment 2: With Degraded Decision Support Machine

In this experiment, we consider 5% and 50% accurate machines.

Normal Set Participants who interact with the 50% accuracy machine and with the 5% accuracy machine experiment a drop in accuracy, suggesting that low accuracy support deteriorates human performance (see left plot in Figure 2). Participants on a 5% accuracy machine showed a slightly negative influence (capacity for correcting mistakes from the machine).

Hard Set For participants interacting with the 50% precision machine, the accuracy does not vary markedly (see the right plot in Figure 2), although the influence is high. For participants interacting with the 5% accuracy machine, there is a marked deterioration in accuracy (see right plot in Figure 2). For these participants, the influence is also high and they tend to be misled by the machine more often than in the normal set.

In Figure 4, we can see that for the participants who label pairs from the Normal Set, the more accurate the machine is, the more useful they find it. However, for the participants who label pairs from the Hard Set, the accuracy or inaccuracy of the machines does not affect as much the perceived usefulness of the machine.

5.3. Experiment 3: With the Variable Decision Support Machine

For every set of pairs, we compare the group of participants who received suggestions from *INC Machine* (accuracies: 5% $\rightarrow$50% $\rightarrow$95%) and *DEC Machine* (accuracies: 95% $\rightarrow$50% $\rightarrow$5%). We distinguish between those participants who were notified every time the machine changed and those who were not.

Normal Set With the *INC Machine*, for both participants with and without notification, accuracy does not vary markedly, and there is no discernible influence by any of the machines. In both cases, confirmation was around 50%. With the *DEC Machine*, for participants with notification, accuracy does not vary markedly, and there is no discernible

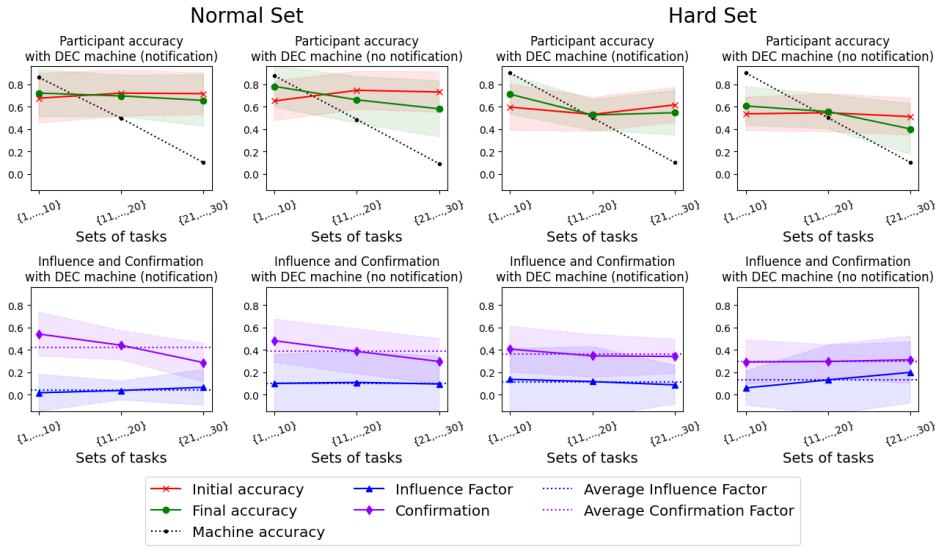


Figure 3. Participant accuracy, Influence and Confirmation factors for DEC machines (with and without notification) with the *Normal Set* (left) and the *Hard Set* (right).

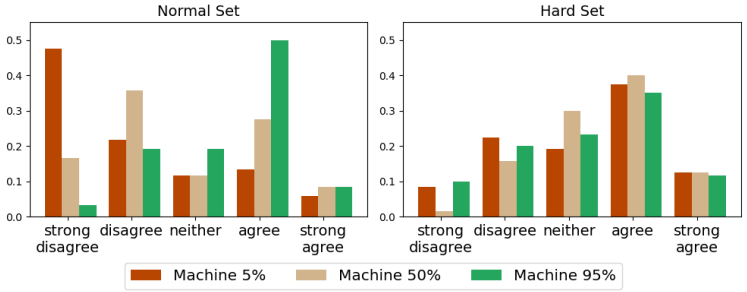


Figure 4. Results obtained from the exit survey from the participants who interacted with some static machine. They were asked whether the suggestions of the machine had been useful to 1. *make a decision when they were not sure about their answer*, 2. *make a decision when some of the images were blurry or the quality was not good*, 3. *make decisions faster*, 4. *confirm their decision when they were certain about it*, 5. *make more accurate decisions*, and 6. *make them feel more confident about the answer*. We show the average of the responses across the six questions.

influence (see plots in first column in Figure 3). For participants without notification, the results show a deterioration in the participant's accuracy, and the mean influence factor noticeably exceeds the influence of participants with notification (see plots in second column in Figure 3).

Hard Set With both *INC Machine* and *DEC Machine*, for both participants with and without notification, accuracy does not vary markedly. In contrast with participants in the *Normal Set*, the influence is now noticeable (see plots in third and fourth columns in Figure 3).

6. Discussion

RQ1 *Does an AI decision support system help improve human performance in a face matching task?*

An accurate machine may improve human performance, but the difficulty of the task might prevent the human from fully exploiting this advantage.

A high accuracy machine improves human performance in both easy and hard tasks, which is aligned with previous works in the literature [6,29]. The influence of this machine is higher when the tasks are harder. However, the improvement and high influence on the Hard Set do not appear to stem from the participant's capacity to recognize the machine's high accuracy since the final questionnaire indicates that inaccurate machines are viewed as equally useful.

RQ2 *Does a low accuracy AI decision support system improve or deteriorate human performance in a face matching task?*

High task difficulty allows an inaccurate machine to induce error more than an accurate machine can induce correctness, probably due to the participants' inability to really grasp how inaccurate the machine is.

For degraded accuracy machines supporting human performance in easy tasks, the degradation of the participant's accuracy is hardly noticeable, suggesting that for a set of easy tasks the participant is able to solve without attending to the machine, as corroborated by the close-to-zero influence values. Observe that the minimal impact of the 95% accuracy machine is partially due to the high confirmation rate (*i.e.*, the machine and user frequently agree, thus reducing the chance of influence), whereas the 5% accuracy machine exhibits a lower confirmation rate yet maintains a marginal influence (*i.e.*, in many tasks, the machine has opposed the participant but did not alter their viewpoint). This suggests that for easy tasks, the participant knows how to solve it well, as it matches the high-accuracy machine, contradicting the low-accuracy one.

However, for difficult tasks, a very low accuracy machine can induce a participant to error very noticeably, even far exceeding the ability of a high-accuracy machine to induce correctness. Our results suggest that for difficult tasks participants tend to be influenced more by the low-accuracy machines than by the accurate machine, while the opposite is true for easy tasks. This may stem from a *projection* of the difficulty experienced by participants. Additionally, the control group's accuracy (no machine) aligns with the 50% accuracy machine. A potential *mirroring* between this machine and its users could explain its significant influence. These phenomena (*self-projection* and *mirroring*) are well established in psychology [30], yet under-explored in human-machine interaction, to the best of our knowledge. Moreover, results from the exit survey highlight that for difficult tasks the participant is not able to perceive a difference in the usefulness of interacting with very accurate or inaccurate machines, which is aligned with some research in the literature [13].

RQ3 *Does a variable AI decision support system improve or deteriorate human performance in a face matching task?*

Automation bias can be induced in a low-performing machine that initially provides accurate support. This can be mitigated with a simple notification that the machine has changed.

For variable machines supporting human performance in easy tasks, the machine influence is barely noticeable except in one case: when a machine initially functions with high precision but gradually loses accuracy. If the participant remains unaware of any change in the machine, they continues to rely on the machine, leading to mistakes. This aligns with the logic of some patterns observed in the literature, where algorithmic aversion is seen to increase when the user sees that the machine fails [15]. We observe something analogous: participants can move from algorithmic appreciation to automation bias after observing machine success. This situation can be prevented by notifying the participant about a machine change without revealing whether it is an improvement or a downgrade. This effect is not observed for variable machines that support human performance in difficult tasks, probably because the participant is not able to clearly identify that the machine is performing accurately at the beginning.

It is notable to observe the distinction between outcomes from the static 50% accuracy machine versus those from machines with varying accuracy levels, increasing or decreasing. While all three machines maintain an average accuracy of 50%, they diverge in how their errors are distributed. Based on our findings, we can deduce that for challenging tasks, randomly distributed errors throughout the interaction promote error camouflage and thus increases the influence factor, compared to machines that accumulate errors at the beginning or at the end of the interaction flow, which have lower influence factors.

7. Conclusions and Future Work

We noted that the difficulty of tasks shapes human-machine interactions in a variety of ways, even when the complexity levels are relatively similar. In our study, there is merely a 10% difference in average human accuracy between *simple* and *complex* tasks. Nevertheless, this small gap appears to be sufficient to notably change the influence of decision support. It is therefore crucial to understand that this challenge relates more to how the participant perceives the task than to their actual skill in solving it. Thus, high difficulty can affect the effectiveness of an accurate machine and can enhance the influence of an inaccurate machine. Conversely, low difficulty can enhance automation bias in the case of variable machines, more specifically those machines that start out accurate (thus eliciting appreciation) and later deteriorate.

In this work we combined the interpretation of the influence factor with the confirmation factor, and that helped us to better understand patterns of interaction. However, the concept of influence allows for many approaches that need to be considered. We have noted that this influence can be negative (aversion), positive (appreciation/automation bias) or neutral (no influence). We observed a tendency towards no influence in easy tasks, and a more erratic tendency towards influence in the case of difficult tasks. This highlights the complexity of measuring influence, necessitating further study on when it is beneficial or detrimental for task completion. We also observed that participants can tell accurate (95%) from inaccurate (5%) decision support when the task is easy, but lose this ability when the task is hard, which makes them more prone to be misled.

A significant constraint identified is the insufficient consideration of real-world, practical concerns. Facial recognition systems are used in complex ethical domains like immigration and law enforcement, where delegating decisions to machines can lead to anonymity, psychological detachment, and invisibility [31,32,33]. These factors may inadvertently promote unethical actions. It is therefore urgent and necessary to extend research on hybrid systems and their corresponding interaction patterns into domains more closely linked to real-world application fields.

Code and Data will be made available with the camera-ready version of this paper.

References

[1] McGee JP, Parasuraman R, Mavor AS, Wickens CD. The future of air traffic control: Human operators and automation. National Academies Press; 1998.

[2] Cummings ML. Automation bias in intelligent time critical decision support systems. In: Decision making in aviation. Routledge; 2017. p. 289-94.

[3] European Union. EU AI Act; 2024. Available from: `https://artificialintelligenceact.eu/the-act/`.

[4] Matias JN. Humans and algorithms work together—so study them together. Nature. 2023;617(7960):248-51.

[5] Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. Journal of Marketing Research. 2019;56(5):809-25.

[6] Logg JM, Minson JA, Moore DA. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes. 2019;151:90-103.

[7] Hou YTY, Jung MF. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW2):1-25.

[8] Mahmud H, Islam AN, Ahmed SI, Smolander K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technological Forecasting and Social Change. 2022;175:121390.

[9] Bigman YE, Gray K. People are averse to machines making moral decisions. Cognition. 2018;181:21-34.

[10] Gogoll J, Uhl M. Rage against the machine: Automation in the moral domain. Journal of Behavioral and Experimental Economics. 2018;74:97-103.

[11] Bonnefon JF, Rahwan I, Shariff A. The moral psychology of Artificial Intelligence. Annual review of psychology. 2024;75(1):653-75.

[12] Önkal D, Goodwin P, Thomson M, Gönül S, Pollock A. The relative influence of advice from human experts and statistical methods on forecast adjustments. Journal of Behavioral Decision Making. 2009;22(4):390-409.

[13] Papenmeier A, Kern D, Hienert D, Kammerer Y, Seifert C. How accurate does it feel?–human perception of different types of classification mistakes. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; 2022. p. 1-13.

[14] Longoni C, Cian L. Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. Journal of Marketing. 2022;86(1):91-108.

[15] Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of experimental psychology: General. 2015;144(1):114.

[16] Dong M, Bocian K. Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions. Journal of Experimental Social Psychology. 2024;111:104584.

[17] Cabitza F, Campagner A, Angius R, Natali C, Reverberi C. AI shall have no dominion: on how to measure technology dominance in AI-supported human decision-making. In: Proceedings of the 2023 CHI conference on human factors in computing systems; 2023. p. 1-20.

[18] Bashkirova A, Krpan D. Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. Computers in Human Behavior: Artificial Humans. 2024;2(1):100066.

[19] Fysh MC, Bindemann M. Human–computer interaction in face matching. Cognitive science. 2018;42(5):1714-32.

[20] Salehi P, Chiou EK, Mancenido M, Mosallanezhad A, Cohen MC, Shah A. Decision Deferral in a Human-AI Joint Face-Matching Task: Effects on Human Performance and Trust. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 65. SAGE Publications; 2021. p. 638-42.

[21] Jeckeln G, Hahn CA, Noyes E, Cavazos JG, O'Toole AJ. Wisdom of the social versus non-social crowd in face identification. British Journal of Psychology. 2018;109(4):724-35.

[22] Ranjan R, Bansal A, Zheng J, Xu H, Gleason J, Lu B, et al. A fast and accurate system for face detection, identification, and verification. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2019;1(2):82-96.

[23] Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, Jackson K, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences. 2018;115(24):6171-6.

[24] Salimzadeh S, He G, Gadiraju U. A missing piece in the puzzle: Considering the role of task complexity in human-ai decision making. In: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization; 2023. p. 215-27.

[25] Majidi F, Khomh F, Li H, Nikanjam A. An Efficient Model Maintenance Approach for MLOps. arXiv preprint arXiv:241204657. 2024.

[26] Faubel L, Woudsma T, Methnani L, Ghezeljhemeidan AG, Buelow F, Schmid K, et al. Towards an MLOps Architecture for XAI in Industrial Applications. arXiv preprint arXiv:230912756. 2023.

[27] Bayram F, Ahmed BS. Towards Trustworthy Machine Learning in Production: An Overview of the Robustness in MLOps Approach. ACM Computing Surveys. 2024.

[28] Renier LA, Mast MS, Bekbergenova A. To err is human, not algorithmic–Robust reactions to erring algorithms. Computers in Human Behavior. 2021;124:106879.

[29] Araujo T, Helberger N, Kruikemeier S, De Vreese CH. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & society. 2020;35(3):611-23.

[30] Waytz A, Mitchell JP. Two mechanisms for simulating other minds: dissociations between mirroring and self-projection. Current Directions in Psychological Science. 2011;20(3):197-200.

[31] Ostermaier A, Uhl M. Spot on for liars! How public scrutiny influences ethical behavior. PloS one. 2017;12(7):e0181682.

[32] Köbis NC, Verschuere B, Bereby-Meyer Y, Rand D, Shalvi S. Intuitive honesty versus dishonesty: Meta-analytic evidence. Perspectives on Psychological Science. 2019;14(5):778-96.

[33] Hancock JT, Guillory J. Deception with technology. The handbook of the psychology of communication technology. 2015:270-89.