UNLOCKING THE POWER OF PARTNERSHIP: HOW HUMANS AND MACHINES CAN WORK TOGETHER TO IMPROVE FACE RECOGNITION

A PREPRINT

P. Jonathon Phillips*
Information Access Division, National Institute of Standards and Technology
100 Bureau Dr., Gaithersburg, MD 20899

Géraldine Jeckeln
School of Behavioral and Brain Sciences, The University of Texas at Dallas, GR41
800 W. Campbell Road, Richardson, TX 75080

Carina A. Hahn
Information Access Division, National Institute of Standards and Technology
100 Bureau Dr., Gaithersburg, MD 20899

Amy N. Yates
Information Access Division, National Institute of Standards and Technology
100 Bureau Dr., Gaithersburg, MD 20899

Peter C. Fontana
Information Access Division, National Institute of Standards and Technology
100 Bureau Dr., Gaithersburg, MD 20899

Alice J. O'Toole
School of Behavioral and Brain Sciences, The University of Texas at Dallas, GR41
800 W. Campbell Road, Richardson, TX 75080

October 6, 2025

ABSTRACT

Human review of consequential decisions by face recognition algorithms creates a "collaborative" human-machine system. Individual differences between people and machines, however, affect whether collaboration improves or degrades accuracy in any given case. We establish the circumstances under which combining human and machine face identification decisions improves accuracy. Using data from expert and non-expert face identifiers, we examined the benefits of human-human and human-machine collaborations. The benefits of collaboration increased as the difference in baseline accuracy between collaborators decreased—following the Proximal Accuracy Rule (PAR). This rule predicted collaborative (fusion) benefit across a wide range of baseline abilities, from people with no training to those with extensive training. Using the PAR, we established a critical fusion zone, where humans are less accurate than the machine, but fusing the two improves system accuracy. This zone was surprisingly large. We implemented "intelligent human-machine fusion" by selecting people with the potential to increase the accuracy of a high-performing machine. Intelligent fusion was more

*To whom correspondence should be addressed. E-mail: jonathon.phillips@nist.gov

accurate than the machine operating alone and more accurate than combining all human and machine judgments. The highest system-wide accuracy achievable with human-only partnerships was found by graph theory. This fully human system approximated the average performance achieved by intelligent human-machine collaboration. However, intelligent human-machine collaboration more effectively minimized the impact of low-performing humans on system-wide accuracy. The results demonstrate a meaningful role for both humans and machines in assuring accurate face identification. This study offers an evidence-based road map for the intelligent use of AI in face identification.

Keywords face identification | decision fusion | wisdom-of-crowds | human-machine collaboration | AI

Significance Statement

Face identifications made by a human working with a computer contribute to decisions in applied and judicial settings. Although two "heads" are usually better than one, individual differences in human and machine performance complicate the decision of when to combine these judgments and when to accept the human or machine decision. We found that the benefits of collaboration for face identification decline as the difference in the ability of the judges (human or machine) increases. Using this rule as a guide for pairing humans with a high-performing machine yielded higher accuracy than always combining human and machine decisions or accepting the machine's (generally) more accurate judgment. AI and human face identification can be improved with intelligent approaches to combining decisions.

1 Introduction

Errors in face identification can have serious consequences for individuals, including unfounded criminal accusations and denial of entry to a country. To minimize errors, facial examiners work in teams; at borders, human inspectors review the results of automated face recognition. Collaborative face identification decisions recall the old adage "two heads are better than one." But, is this always true? And, what happens if one of the two "heads" is a computer? Multiple studies indicate that face identification accuracy increases when two humans collaborate in making face identification decisions [1, 2, 3, 4, 5, 6]. We consider the juxtaposition between the consistent face identification fusion benefits observed in the *general* case and the variability of these benefits on a case-by-case basis. Whereas the overall benefit of fusion for face identification replicates across studies, within each study, not every human partnership increases identification accuracy. Fusing the judgments of individual people can increase performance, decrease it, or leave it unchanged [7, 5, 8]. In consequential applications, it is critical to be able to predict the likelihood of success for individual fusion cases.

In moving from the consistency of fusion benefits in general, to the case of specific individuals, one factor implicated in fusion success is the relative difference in ability between the two individuals to be fused. This has been studied for human collaboration with simple visual perception tasks [9, 10, 11] and medical diagnoses [12]. The results converge on a remarkably simple rubric for fusing human decisions: Human judgments should be combined only when the baseline accuracy of the observers is similar. We term this the *proximal accuracy rule (PAR)*. The first goal of this study was to determine whether the PAR governs human partnerships for face identification.

More broadly, understanding the mechanics of decision fusion is also important for combining human and machine judgments. Machines are now an integral part of identification systems in applied scenarios. When the decisions of a human and a machine differ, it is critical to determine whether/how to combine the two judgments. Similar to the fusion of human decisions, combining human and machine judgments of face identification increases accuracy [13, 5]. But this does not occur in every individual case. Moreover, differences in the ability of the human and machine have not been investigated as a factor in fusion success—except in the case where participants had *a priori* knowledge about the baseline accuracy of the machine [cf., 7]. The superiority of automated face recognition over humans in some cases [e.g., 5], combined with societal distrust of using an AI system without human oversight [e.g., 14], poses a challenge for assuring accurate face identification in consequential settings.

We tested whether the PAR could be used to predict fusion success for individual human-human and human-machine partnerships. At a systems level, we tested whether intelligent fusion, guided by the PAR, could increase face identification accuracy across a group of humans working with a high-performing machine. For human-only partnerships, we optimized dyad pairing using graph matching theory and compared this human-only system to a system of intelligent human-machine partnerships [15, 16]. The findings point to a meaningful role for both humans and machines in assuring accurate face identification.

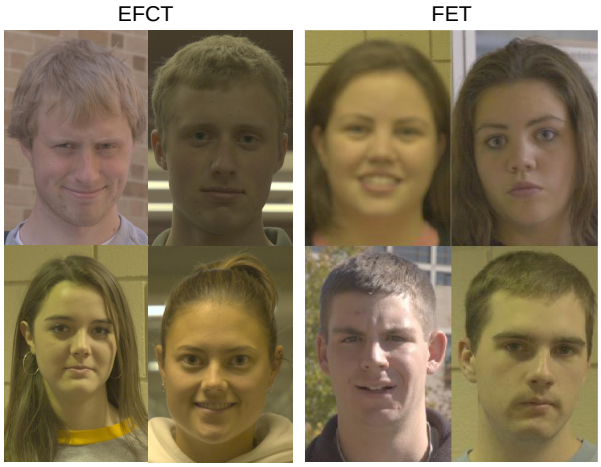

Figure 1: Example image-pairs from the EFCT (left) and the FET (right). The top row shows same-identity face image pairs and the bottom row shows different-identity pairs (faces cropped from full image).

Results

Two face identification tests were used to address these questions: the Expertise in Facial Comparison Test (EFCT) [8] and the Facial Expertise Test (FET) [5] (see Figure 1 for example face image pairs). Both tests included human participants across a wide range of skill levels from untrained university students to highly trained professional forensic face examiners. Participants viewed pairs of images and rated their certainty that the images in the pair showed the same person versus different people. For the machine used in each task, identification "certainty" was measured as the similarity between machine-generated representations of the images. Fusion was implemented by averaging the certainty ratings for human-human and human-machine partnerships. Face identification accuracy was measured as area under the Receiver Operating Characteristic Curve (AUC) (1 indicates perfect performance, 0.5 indicates random performance). *Fusion benefit* was defined as the difference between a pair of collaborators' fused performance (human-human or human-machine) and the performance of the more accurate person/machine in the pair.

Proximal Accuracy Rule predicts fusion benefits for human partnerships

Dyads consisted of all possible pairs of individual humans. For each pair, certainty ratings on each test item were averaged and a fused AUC was calculated. The benefit of human-human fusion decreased as the difference in accuracy between the two participants increased (Fig. 2, top). This was true for both the EFCT and FET. Consistent with the PAR, there was a strong negative correlation between fusion benefit and the difference in baseline accuracy of the paired individuals (EFCT: $r(2626) = -0.7385$, $p < .001$, 95% CI [-0.7554, -0.7207]; FET: $r(17576) = -0.7052$, $p < .001$, 95% CI [-0.7126, -0.6977]).

The PAR predicted fusion benefits across the wide range of accuracy represented in this pool of participants (colored dots in Fig. 2, in top row indicate the accuracy of the better performer in each dyad). The figure shows that fusion benefits are robust even when the better performer in the dyad performs poorly.

Further analyses revealed that the PAR is independent of the performance level of the worst (best) performer within each dyad. Specifically, for human-human dyads classified into three performance categories (low, medium, high) based on the performance of the worst (best) performer, the fusion benefit increased as the baseline accuracy difference between the paired individuals decreased (see Supporting Information).

Proximal Accuracy Rule predicts fusion benefits for human-machine partnerships

Dyads consisted of the machine paired with each human. The benefit of human-machine fusion decreased as the difference in accuracy between the human and machine increased (Fig. 2, middle). This was the case for both tests. Consistent with the PAR, there was a strong negative correlation between fusion benefit and the difference in baseline accuracy of the paired human and machine (EFCT: human partnered with VGG-Face [17], $r(71) = -0.8996$, $p < .001$, 95% CI [-0.9359, -0.8442]; FET : human partnered with A2017b [18], $r(182) = -0.7915$, $p < .001$, 95% CI

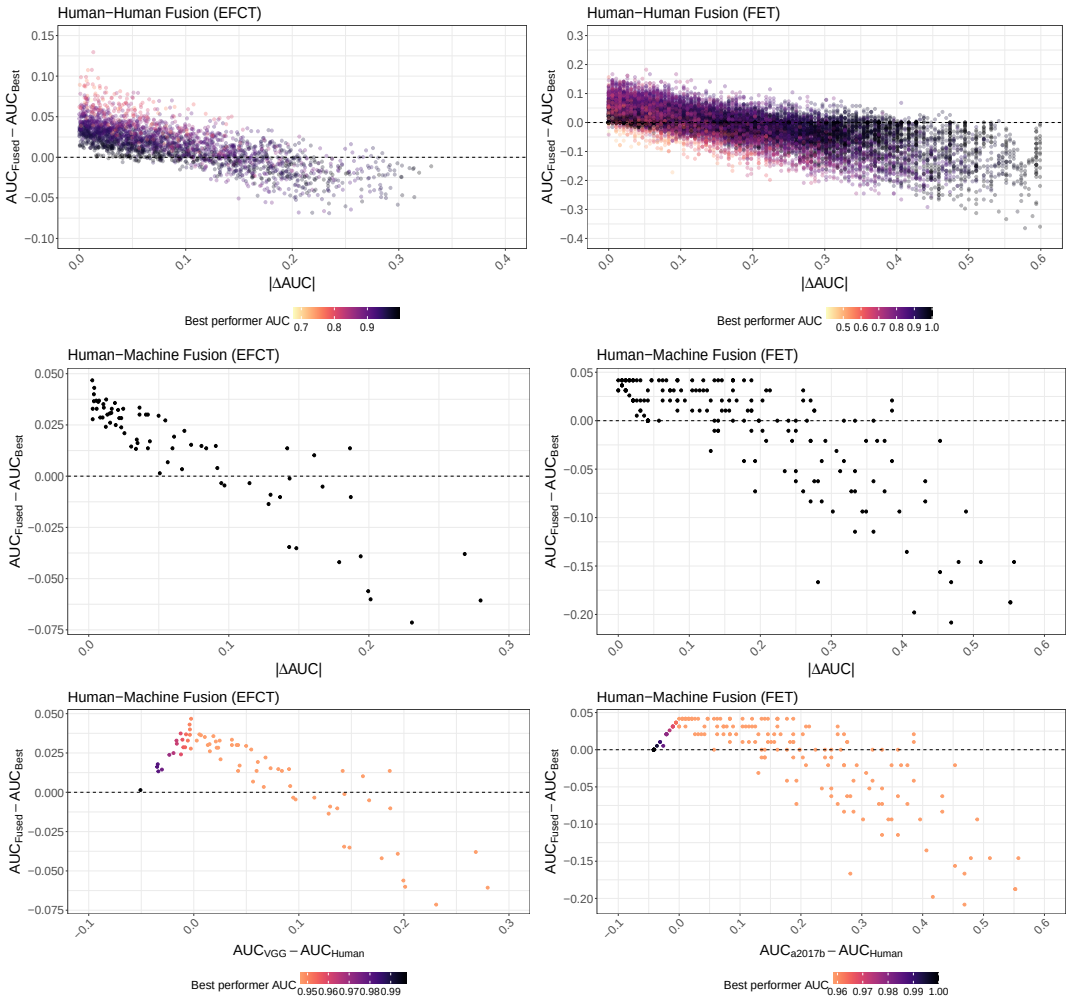

Figure 2: **Fusion benefits for EFCT (left) and FET (right) data.** Top row. Human-human fusion benefit ($AUC_{fused} - AUC_{best}$) of each dyad (dot) is plotted as a function of the absolute difference in accuracy between dyad members ($|\Delta$AUC$|$). The dots represent all possible human-human dyads and the color indicates the baseline performance of the better performer in each dyad. Fusion benefit decreases with increased difference in the baseline performance of the dyad members for the human-human dyads. Middle row. Human-machine fusion benefit of each dyad (dot) plotted against the absolute difference in accuracy between the human and machine (VGG-Face for EFCT and A2017b for FET) ($|\Delta$AUC$|$). Note that machine performance is a constant for each test. Bottom row. Fusion benefit as a function of how much better (worse) the machine performed than the human performed. Fusion is more beneficial than accepting the better performer's decision (dots below the line) for dyads that fall above the horizontal lines.

[-0.84, -0.7305]). Note that baseline performance for the machines is constant (VGG-Face, $AUC = 0.946$; A2017b, $AUC = 0.9583$).

When the human was less accurate than the machine, combining the two decisions remained beneficial up to a surprisingly large machine advantage. This is illustrated in Fig. 2 (bottom), which shows the benefit of fusion as a function of how much more (less) accurate the human was than the machine. Specifically, human-machine fusion was beneficial up to a machine advantage in AUC of ≈ 0.10 for the EFCT ($AUC_{VGG-Face} > AUC_{human}$) and AUC of ≈ 0.20 for the FET ($AUC_{A2017b} > AUC_{human}$). These numbers define a *critical fusion difference* zone of human ability below a machine's baseline accuracy. Participants with scores in this zone can improve a machine's decision. In practical terms, the critical fusion difference indicates, for example, that a human with a baseline accuracy of $AUC = 0.85$ can still add to the accuracy of VGG-Face ($AUC = 0.9467$); a human with a baseline accuracy of $AUC = 0.75$ can still add to the accuracy of A2017b ($AUC = 0.9583$).

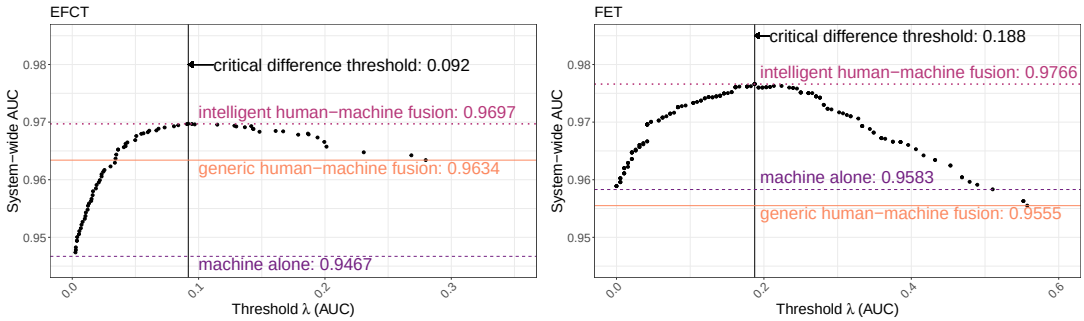

Figure 3: **Intelligent Human-Machine System-wide Partnership**. The intelligent fusion rule analysis is presented for the EFCT (left) and FET (right) data sets. The graphs show the system-wide AUC (vertical axis) as a function of the threshold λ (horizontal axis). Below the threshold λ, humans are fused with the machine; above the threshold λ, the machine's decision is accepted. The critical difference threshold (optimal threshold) is marked by the vertical line. To allow for comparisons among different methods for implementing decisions, horizontal lines show system-wide AUCs for machine operating alone (dashed purple line), generic human-machine fusion (solid orange line), and intelligent human-machine fusion (dotted pink line). The system-wide AUC for humans alone is 0.885 for the EFCT and 0.796 for the FET (not on the graphs).

When the human was more accurate than the machine (Fig. 2 bottom, negative values on the $x - axis$), human-machine fusion was always beneficial. The uniformity of this latter finding is due to the high baseline performance of both machines on the face identification tests (VGG-Face for EFCT, $AUC = 0.9467$, A2017b for FET, $AUC = 0.9583$). This level of machine performance places a ceiling on the human advantage over the machine, which falls within the critical fusion zone. In other words, human performance would have to exceed 1.0 (perfect) for the machine not to be beneficial.

In summary, the PAR accurately predicts fusion success for both human-human and human-machine dyads. Smaller differences in baseline ability yield larger fusion benefits. Given a large enough disparity in baseline performance between the two participants, fusion can degrade dyad performance to a level below that of the more accurate participant. However, the relatively large critical fusion zones indicate that human-machine fusion can be beneficial even when the baseline accuracy of a human is substantially lower than that of the machine.

Intelligent human-machine fusion improves system-wide accuracy

With knowledge of when to fuse the face identification decisions of individual humans with those of a machine, it is possible to explore how much face identification accuracy can be improved at a system-wide level by implementing the PAR. We define a *system* as an entire group of humans each partnered with the same machine. System performance was calculated as a function of the difference in human and machine baseline performance permitted for fusion of the two judgments. Beyond this difference the identification decision defaulted to the machine judgment. Specifically, we varied the $|\Delta\text{AUC}|$ threshold (λ), while implementing a fusion rule of the form:

$$decision = \begin{cases} \text{fuse human and machine,} & \text{if } |\Delta AUC| \leq \lambda \\ \text{accept machine decision,} & \text{otherwise} \end{cases}$$

At each λ, we computed system-wide performance as the average of AUCs for the fused dyads and "dyads" that defaulted to the machine's decision. We varied $|\Delta AUC|$ from levels where no human is fused with the machine (machine decision accepted) to values where all humans are fused with the machine.

Figure 3 shows how system-wide performance changes as λ is varied for human-machine dyads in the EFCT (n = 73) (left) and FET (n = 184)(right). Both tests show a similar pattern of results. When no human is fused with the machine, the machine's baseline performance provides a floor for system-wide performance (purple horizontal line). As the threshold is increased, system-wide performance increases steadily until the threshold reaches the critical fusion value (EFCT λ = 0.092; FET λ = 0.188) (black vertical line). At this critical fusion threshold, the system of humans partnering with a machine achieves its highest accuracy—we refer to this as *intelligent fusion*. Performance declines as human-machine fusions are included that are outside of the critical difference between the human and machine.

One difference between the EFCT and FET simulation results is the ordering of accuracy when all humans are fused with the machine versus when the machine judgment is always accepted. For the EFCT system, the former is superior

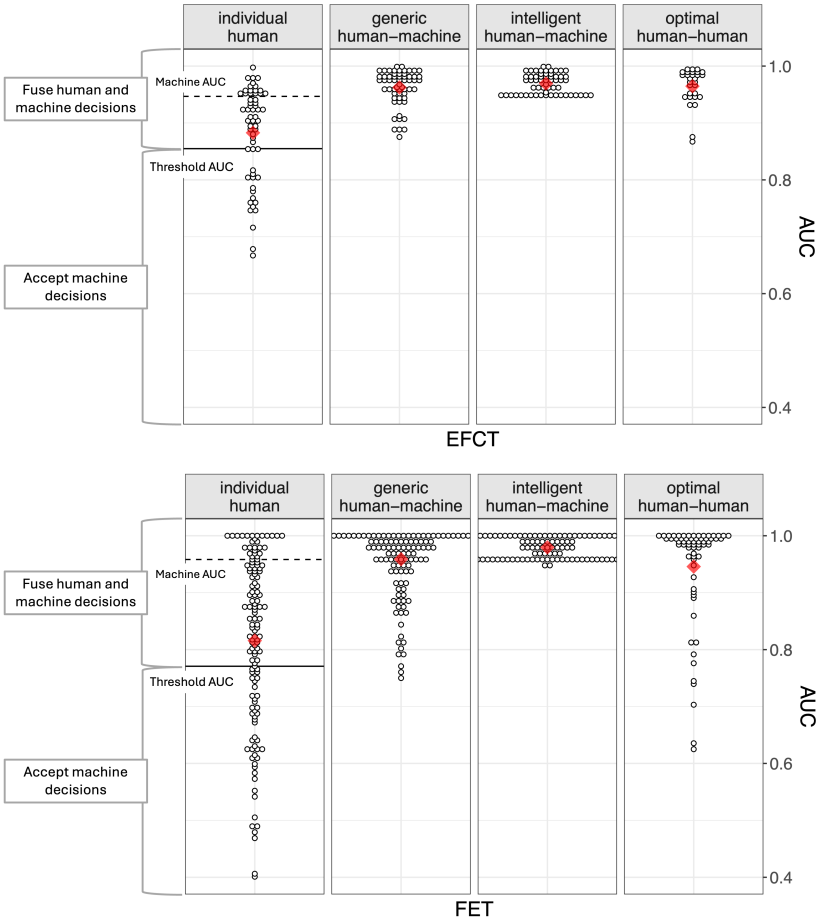

Figure 4: **Compare different fusion methods.** The comparison for the EFCT (top) and FET (bottom). Each graph summarizes performance for four conditions (columns from left to right): individual humans working alone, generic human-machine fusion, intelligent human-machine fusion, and optimal human-human fusion. For individual humans, each circle is the AUC of an individual and the red diamond marks the median human accuracy. The dashed line shows the machine AUC. For intelligent fusion, humans above the solid line labeled 'Threshold AUC' should be fused with the machine and those below the line should not be fused with the machine. The critical fusion zone corresponds to the AUCs between the machine AUC (dashed line) and the threshold AUC (solid line). In generic human-machine fusion, every human is fused with the machine. In this panel, each circle reports the AUC for a human-machine dyad, and the red diamond is the median AUC for the human-machine dyads. In the intelligent fusion panel, each circle is either the AUC of a human-machine dyad or the machine alone. This depends on the AUC of the individual in the individual human panel. The red diamond is the median of the circles in this panel. The optimal human-human panel shows the AUC of the optimal human partners, and the red diamond is the median of these partners.

to the latter. For FET, the machine operating alone is more accurate than fusing all humans with the machine. This difference is a natural consequence of two things: the greater difficulty of the FET over the EFCT and the superiority of A2017b (FET partner) over VGG (EFCT partner). The divergence of ordering here illustrates that the average and range of individual human performance, relative to machine performance, must be taken into account when fusing human and machine decisions at a system-wide level. This is consistent with the general principle of the PAR.

Figure 4 illustrates the importance of individual differences in human performance for fusion success at a system level. Participants on the EFCT and FET vary widely in accuracy, with numerous individuals performing quite poorly. For the more challenging test (FET), the distribution of human scores has a particularly extended tail (Fig. 4). When all humans are fused with the machine (*generic human-machine fusion*), the system's average performance increases substantially over the performance of individuals and the tail of the distribution shrinks (Fig. 4). Generic fusion is

superior to individual performance (Bonferroni corrected Wilcoxon Signed Rank Paired Test: EFCT, $W = 0$, $p \approx 0$; FET, $W = 0$, $p \approx 0$).

Intelligent human-machine fusion using the PAR shrinks the distribution tails further, without changing the average performance substantially (Fig. 4). Intelligent fusion is superior to both generic fusion (Bonferroni corrected Wilcoxon Signed Rank Paired Test: EFCT, $W = 30$, $p = 0.009$; FET: $W = 103$, $p \approx 0$) and individual performance (Bonferroni corrected Wilcoxon Signed Rank Paired Test: EFCT, $W = 0$, $p \approx 0$; FET: $W = 0$, $p \approx 0$).

To summarize, maximal system performance was achieved by combining the machine's decisions with the decisions of humans who perform at or above the level of the machine, and humans less accurate than the machine, but within the critical fusion zone.

Optimal human partnering improves system-wide face identification accuracy

Intelligent human-machine fusion shows that there are cases when face identification decisions are best left to the machine. Here, we determined the best accuracy attainable by a system of humans partners. The goal was to find the set of human dyads that would yield the highest *system accuracy*, defined as the average AUC for the selected dyads. In this system, a single human can serve in only one dyad. From N people, we generate $N/2$ dyads when N is even, and $(N-1)/2$ dyads when N is odd.

To find optimal dyads, we turn to a branch of mathematics called graph theory. To solve for the optimal set of dyads, we formulated the problem as a weighted graph matching problem [15, 16], (see Methods). In this formulation, the fused AUCs for all possible dyads serve as input to the graph matching algorithm. The optimal set of partners is output. The performance of these optimal dyads is illustrated in Fig. 4. Performance was greater for optimal human-human partnering than for individual humans [Mann-Whitney U test (unpaired): EFCT, $U = 198$, $p \approx 0$; FET: $U = 2670$, $p \approx 0$].

Next, we assessed whether optimal human partnering is substantially better than randomly selecting partners. To test this, we generated 100 systems consisting of random dyads, enforcing the rule that a person can only be in one dyad. The mean was computed for each of the random system. There was no overlap between the optimal system mean and the accuracy distribution of the means for the random systems. For the EFCT, the optimal system achieved an accuracy of 0.973. Random system accuracy (Mean = 0.9463; $sd = 0.0026$) differed from the optimal system accuracy by 10.40 standard deviations. For the FET, the optimal system achieved an accuracy of 0.962. Random system accuracy (Mean $= 0.874$; $sd = 0.0040$) differed from the optimal system accuracy by 22.15 standard deviations. The optimal system was far superior to randomly partnering people.

Human, machine, both?

So far, we have studied fusion for two groups of people from two datasets with each group having a wide range of abilities (face experts to untrained students). We found that intelligent human-machine fusion and optimal human-human partnering both yield higher accuracy than individuals operating alone. Both also yield more accurate performance than their non-selective fusion counterparts (generic fusion, random dyad fusion). The accuracy of the machine and the shape of the human performance distribution, however, must be considered in deciding whether the inclusion of a machine in the system is beneficial.

Performance distributions for intelligent human-machine fusion and optimal human-human fusion appear in Fig. 4. The average accuracy for optimal human dyads approximates the average accuracy achieved by intelligent human-machine fusion, but the shape of the distributions differs. The tails of the human-only systems are extended relative to those of intelligent human-machine fusion. This is particularly evident for the more difficult FET, which has a wider range of individual human performance than the EFCT. For the FET, the human-machine system was more accurate than the optimal human-human system (Mann-Whitney U test, $U = 6760$, $p = 0.00235$). For the less challenging EFCT, the two fusion methods yielded comparable performance ($U = 1170$, $p = 0.347$).

The primary lesson to be learned from comparing a fully human system to a human-machine system is that individual differences in human performance can be a liability in the fully human system. If the range of human performance is wide, the benefits of fusion can be diluted by combining the decisions of individuals with accuracy differences that exceed the critical fusion threshold. Intelligent human-machine fusion has the advantage of minimizing the impact of low performing humans on system-wide performance.

Discussion

In an era where humans and AI partner to perform important and consequential tasks, it can be difficult to know what to do when a human and machine disagree. In cases of conflict between two decisions, it may seem intuitively reasonable to simply choose the better system (human or machine). Or, if human review is a priority, select the combined human-machine decision. In both cases, this would be a mistake. Collaborative decision-making offers a better alternative, when it is applied intelligently.

Investigating the challenges of partnering people with each other and with face recognition technology, we built on lessons learned from the decision-making literature about how best to combine human decisions [9, 11, 12]. We found that the Proximal Accuracy Rule, which predicts fusion benefits in other diverse human decision tasks [9, 11, 12], generalizes seamlessly to face identification. This was true for human-human collaborations and for human-machine collaborations. The applicability of the PAR across multiple diverse tasks suggests that effective human oversight of machine-generated decisions can come from leveraging knowledge about the baseline abilities of the individuals and machines to-be-fused. For face identification, combining the decisions of two judges (human or machine) is most advantageous when the baseline abilities of the two are similar. Combining judgments amounts to simply summing the decision certainties. This is consistent with longstanding theory in combining pattern classifier decisions [19].

An important feature of the PAR is that it applies equally to people with varying levels of face identification training and ability. Fusion success can be predicted for participants ranging from university students (no experience/poor ability) to professional forensic face examiners (extensive professional training/high ability). It is perhaps worth pausing to state the obvious; the most accurate face identification will come from employing people and/or machines with the highest levels of skill. In judicial settings, for example, it is possible to select individuals with the highest level of skill and to assure adequate training. Professional forensic face examiners, reviewers, and super-recognizers surpass novices on face identification tests. Even in this best case scenario, there is a wide range of individual performance [e.g., 5, 8]. In other consequential cases, for example, passport examiners at border crossings, it may be more difficult to assure uniformly high levels of skill and adequate training. The applicability of the PAR across skill levels makes it possible to improve face identification in both types of scenarios.

In examining fusion benefits across a system of human collaborators, graph theory provided a novel way of optimizing partnerships across system-wide human resources (i.e., people). Graph theory has been applied to problems in diverse domains, including in physics, chemistry, and linguistics. As applied here, weighted graph matching takes into account individual differences in human skill and dyad fusion benefits across a large group of people. It ultimately enabled a computationally efficient solution to this combinatorial problem. The dyad pairing solution provided by graph theory yielded performance far better than that achievable by randomly pairing individuals.

An ongoing challenge for collaboration going forward is that face identification technology continues to improve. Machine performance now falls within the range of human accuracy, often near the top of the human distribution [5, 20, 21, 22]. This is true even for highly challenging tasks such as discriminating the faces of identical twins [20]. Given a machine that performs well relative to humans on a give face identification task, many people will perform less accurately than the machine. We showed that input from a human judge can be beneficial up to a surprisingly large machine advantage. The critical fusion threshold we found for both the EFCT and FET data sets indicates that we should not summarily discard human input from individuals who have less ability than a machine. This would produce less accurate face identification in many cases. Instead, systematic determination of the critical fusion zone provides a better estimate of whether a person with less ability than a machine will add to decision accuracy.

In applied face identification scenarios of consequence, system-wide performance is only as good as its weakest link. The ability of intelligent human-machine fusion to shrink the tail of the individual performance distribution, therefore, has value in forensic applications. Studies with both face experts and novices commonly show a broad range of individual differences in accuracy [23, 24, 25, 5, 8, 26]. Occasionally, the lowest-performing experts are no more accurate than some novices [5, 8]. The high performance of machines, relative to most humans, can serve to put a floor on the performance distribution.

More broadly, fusion success requires that the computational strategies of the judges to-be-fused are both well-grounded and divergent [cf. 13]. By well-grounded, we mean that they are valid or useful. For example, trying to identify faces using only the eyes is a valid, albeit sub-optimal, face identification strategy. More valid strategies lead to more accurate face identification. By divergent, we mean strategies that are not identical. For example, one person might try to identify a face using only the eyes and a second person might use only the top of the face. These are divergent strategies. The present results demonstrate that strategies with proximal levels of validity lead to fusion benefits. And, of practical importance, the critical fusion threshold can be established empirically for any given scenario.

Understanding the role of divergence in fusion is more challenging. Humans vary in both in overall ability and in their approach to face recognition. Human face recognition ability naturally spans levels from super-recognizers to developmental prosopagnosics [27]. It is not clear how the strategies of the best identifiers differ from those of less skilled individuals [28, 29, 30]. Despite far better performance, super-recognizers and developmental prosopagnosics use the same critical features for face identification as normal individuals [31], possibly indicating that the best face identification comes from fully exploiting the most useful identity information in a face.

Differences in human ability are complemented by differing qualitative patterns of face recognition—the distinct and unique ways that individuals perceive and identify faces. These qualitative differences give rise to variability in the pattern of errors across a set of faces. Individual experience with particular sets of familiar faces may partially account for different patterns of errors [32]. Notably, face recognition is more accurate for faces that resemble people we know [32]. Idiosyncratic divergence in the set of people we know may lead to variability in the errors we make. This "islands of expertise" hypothesis suggests that personal experience may increase face processing expertise in regions of our face space [21, 33] around the representations of familiar others [32]. Going forward, the PAR might prove beneficial in selectively combining individuals across islands of expertise to potentially benefit from divergent strategies of recognition.

Returning to the original question of whether two heads are better than one for face identification, we conclude that this depends on the difference in baseline ability of the two performers. This is true even when one of the "heads" is a computer. As machines contribute to impactful tasks, methods for combining human and machine decisions need to be based on theories that are valid for both humans and machines. In our case, we started with the psychological decision theory of humans, which explicitly accounts for the properties of how humans arrive at decisions. Our findings demonstrate that the PAR and fusion properties naturally extend to human-machine collaborations. The PAR gave us the critical fusion zone, which shows that combining humans and machines is valuable even when human ability is significantly less than that of machines.

Methods

Data sets

Human-human and human-machine fusion was evaluated with a face-identity matching task. To that end, we used existing data from the Expertise in Facial Comparison Test (EFCT) [8] and the Facial Expertise Test (FET) [5]). For the EFCT, we used human responses from the condition in which stimuli were presented upright for 30 seconds. Both datasets include participants who span a broad range of ability, from experts trained for face identification to untrained university students.

EFCT. The EFCT included 84 pairs of face images (42 same-identity and 42 different identity pairs). The task was to judge whether the image pairs portray the same identity or different identities (see example in Figure 1 left). Participants judged the similarity of the two faces using a 5-point scale (1: sure they are the same person; 5: sure they are different people). Participants ($n = 73$) included 27 forensic face examiners, 32 university students, and 14 participants attending the Facial Identification Scientific Working Group (FISWG) [8]. The FISWG participants were selected as controls due to their interest in the topic and motivation to perform well. These participants were not trained in face identification.

Additionally, we used a deep convolutional neural network (DCNN) (VGG-face [17]) to obtain similarity measures for each EFCT face-image pair. Similarity was computed following methods described in [34].

FET. The FET included 20 challenging face-image pairs (12 same-identity and 8 different-identity pairs) [5] (see example in Figure 1 right). Again, the task was to determine whether the two images portray the same or different identities. For the human data, participants responded using a 7-point scale (-3: high confidence that the pair showed different people; +3: high confidence that the pair showed the same person). The human participants ($n = 184$) included 57 forensic facial examiners, 30 facial reviewers, 13 super-recognizers, 53 fingerprint examiners, and 31 students (controls) [5]. Fingerprint examiners were included as controls who were forensically trained on fingerprint identification, but not on face identification. For the machine input, we used the most accurate of the four DCNNs tested in previous work [5], A2017b [18].

All analyses and simulations were performed for the data from each test (EFCT and FET), separately.

Data Availability

For the FET, deidentified data for facial examiners and reviewers, superrecognizers, and fingerprint examiners can be obtained by signing a data transfer agreement with the NIST. Data for the students and algorithms are available as

supplemental material in [5]. For the ECFT, permission to release deidentified data for the participant was not obtained. The FET and ECFT images are available by license from the University of Notre Dame.

Face Identification and Fusion Benefit Metrics

Baseline Face Identification Accuracy

Accuracy was measured as the area under the Receiver Operating Characteristic Curve (AUC) computed from the distributions of face-matching responses (human observers) or similarity scores (DCNN) for same-identity pairs and different-identity pairs. We evaluated baseline human accuracy using the participants' independent responses for each image pair. Similarly, we evaluated the baseline DCNN accuracy using the DCNN-based similarity scores for each image pair.

Fusion: human-human and human-machine

Fusion was achieved by combining the responses from two human observers (human-human fusion) or the responses from a human and a DCNN (human-machine fusion).

For human-human fusion, human observers were assigned to dyads, using methods that varied by the analysis (e.g., all possible combinations, random). Next, for each dyad, we averaged the two humans' responses for each image pair. These averaged responses were then used to compute the fused accuracy (AUC_{fused}).

For human-machine fusion, human observers were paired with a DCNN to create dyads, also using methods that varied by the analysis (e.g., generic, intelligent). For the EFCT data, participants were paired with VGG Face [17]. For each face-image pair (i), the DCNN-based similarity score (s_i) was scaled to the distribution of responses from all human participants in the EFCT data set as follows:

$$\widehat{s}_i = \left(\frac{s_i - \mu_A}{\sigma_A} \right) + \mu_H.$$

where μ_A and σ_A denote the mean and standard deviation of the original machine similarity scores, and μ_H denotes the mean of human responses. For each human-machine dyad, on each face-image pair, we averaged the DCNN's scaled similarity score ($\widehat{s}_i$) and the human's response.

For the FET data, participants partnered with A2017b [18] and the averaged responses were obtained from a previous study [cf., 5]. For each human-machine dyad, fused accuracy (AUC_{fused}) was computed using the averaged responses for each image pair.

Absolute difference in baseline accuracy between the two participants (human-human or human-machine) in each dyad, $|\Delta\text{AUC}|$, was computed as:

$$|AUC_{performer1} - AUC_{performer2}|$$

Fusion benefit was measured by subtracting the accuracy of dyad's best performer (AUC_{best}) from the dyad's fused accuracy (AUC_{fused}):

$$AUC_{fused} - AUC_{best}$$

Optimal Human Dyads for Face Identification

Graph Theory

The maximum weighted matching was computed for a complete graph with N vertices, where N is the number of humans. Each vertex v_i corresponds to the i^{th} person. The weight of the edge between v_i and v_j is the AUC of fusing the i^{th} and j^{th} person. We solved this maximum weighted matching problem using the python NetworkX package [35].

Acknowledgments

Research at the University of Texas at Dallas was funded by The National Institute of Standards and Technology, Grant 70NANB21H109 & 70NANB22H150 to A.OT.

References

[1] Jacqueline G. Cavazos, Géraldine Jeckeln, and Alice J. O'Toole. Collaboration to improve cross-race face identification: Wisdom of the multi-racial crowd? *British Journal of Psychology*, 114:838–853, 2023.

[2] Andrew J Dowsett and A Mike Burton. Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, 106(3):433–445, 2015.

[3] Géraldine Jeckeln, Carina A. Hahn, Eilidh Noyes, Jacqueline G. Cavazos, and Alice J. O'Toole. Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology*, 109:724–735, 2018. ISSN 20448295. doi:10.1111/bjop.12291.

[4] Tarryn Balsdon, Stephanie Summersby, Richard I Kemp, and David White. Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, 3:1–13, 2018.

[5] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[6] David White, A Mike Burton, Richard I Kemp, and Rob Jenkins. Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology*, 27(6):769–777, 2013.

[7] Daniel J. Carragher and Peter J. B. Hancock. Simulated automated facial recognition systems as decision-aids in forensic face matching tasks. *Journal of Experimental Psychology. General*, 152, 2022.

[8] David White, P. Jonathon Phillips, Carina A. Hahn, Mathew Q. Hill, and Alice J. O'Toole. Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B*, 282, 2015.

[9] Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

[10] Dan Bang, Riccardo Fusaroli, Kristian Tylén, Karsten Olsen, Peter E Latham, Jennifer YF Lau, Andreas Roepstorff, Geraint Rees, Chris D Frith, and Bahador Bahrami. Does interaction matter? testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition*, 26:13–23, 2014.

[11] Asher Koriat. When are two heads better than one and why? *Science*, 336(6079):360–362, 2012.

[12] Ralf H. J. M. Kurvers, Stefan M. Herzog, Ralph Hertwig, Jens Krause, Patricia A. Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek, and Max Wolf. Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113:8777–8782, 2016. ISSN 0027-8424. doi:10.1073/pnas.1601827113.

[13] Alice J O'Toole, Hervé Abdi, Fang Jiang, and P Jonathon Phillips. Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1149–1155, 2007.

[14] Johann Laux. Institutionalised distrust and human oversight of artificial intelligence: Toward a democratic design of ai governance under the european union ai act. *AI & Soc*, 39:2853–2866, 2024.

[15] Jack Edmonds. Maximum matching and a polyhedron with (0, 1) vertices. *J. Res. Natl. Bureau Standards B*, 69, 1965.

[16] Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.

[17] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision*, 2015.

[18] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. In *arXiv preprint arXiv:1703.09507*, 2017.

[19] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. doi:10.1109/34.667881.

[20] Connor J Parde, Virginia E Strehle, Vivekjyoti Banerjee, Ying Hu, Jacqueline G Cavazos, Carlos D Castillo, and Alice J O'Toole. Twin identification over viewpoint change: A deep convolutional neural network surpasses humans. *ACM Transactions on Applied Perception*, 20(3):1–15, 2023.

[21] Alice J O'Toole and Carlos D Castillo. Face recognition by humans and machines: three fundamental advances from deep learning. *Annual Review of Vision Science*, 7(1):543–570, 2021.

[22] Géraldine Jeckeln, Selin Yavuzcan, Kate A Marquis, Prajay S Mehta, Amy N Yates, P Jonathon Phillips, and Alice J O'Toole. Designing cross-race tests for forensic facial examiners, super-recognizers, and face recognition algorithms. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2024.

[23] Josh P Davis, Karen Lander, Ray Evans, and Ashok Jansari. Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology*, 30:827–840, 2016.

[24] Kristin Norell, Klas Brorsson Läthén, Peter Bergström, Allyson Rice, Vaidehi Natu, and Alice J. O'Toole. The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences*, 60(2): 331–340, 2015.

[25] Alice Towler, David White, and Richard I Kemp. Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, 23(1):47, 2017.

[26] David White, Alice Towler, and Richard I Kemp. Understanding professional expertise in unfamiliar face matching. In M. Bindemann, editor, *Forensic Face Matching: Research and Practice*, pages 62–88. Oxford University Press, 2021.

[27] Richard Russell, Brad Duchaine, and Ken Nakayama. Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, 16(2):252–257, 2009.

[28] Eilidh Noyes, P. Jonathon Phillips, and Alice J. O'Toole. What is a super-recogniser? In M. Bindermann and A. M. Megreya, editors, *Face Processing: Systems, Disorders, and Cultural Differences*. Nova, New York, NY, USA, 2017.

[29] Meike Ramon, Anna K Bobak, and David White. Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, 110(3):461–479, 2019.

[30] Andrew W Young and Eilidh Noyes. We need to talk about super-recognizers Invited commentary on: Ramon, M., Bobak, A. K., & White, D. Super-recognizers: From the lab to the world and back again. British Journal of Psychology. *British Journal of Psychology*, 110(3):492–494, 2019.

[31] Naphtali Abudarham, Sarah Bate, Brad Duchaine, and Galit Yovel. Developmental prosopagnosics and super recognizers rely on the same facial features used by individuals with normal face recognition abilities for face identification. *Neuropsychologia*, 160:107963, 2021.

[32] Peter JB Hancock. Familiar faces as islands of expertise. *Cognition*, 214:104765, 2021.

[33] T. Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology*, 43:161 – 204, 1991.

[34] P. Jonathon Phillips. A cross benchmark assessment of deep convolutional neural networks. In *12th IEEE Conference on Automatic Face and Gesture Recognition*, 2017.

[35] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

Supporting Information

Experiment details

Proximal accuracy rule for human-human dyads

For each dataset (EFCT and FET), we tested whether the PAR is independent of the level of performance exhibited by the worst (best) performing member within each dyad. First, dyads were ranked based on the accuracy (AUC) of their worst (best) performer. Second, the range in accuracy (based on the worst/best performer) occupied by all the dyads in each data set was divided into three levels (low, medium, high). For each level, we computed the correlation between the benefit of fusion and the difference in baseline accuracy of the paired individuals. Correlation results based on the "worst performer" and "best performer" of each dyad are shown in Table 1 and Table 2, respectively. Overall, the results indicated that the PAR applied to dyads across all levels of performance (low, medium, high). Specifically, the benefit of fusion increased as the difference in baseline accuracy of the paired individuals decreased. The results are shown in Figure 5 for the dyads' "worst performer" and Figure 6 for the dyads' "best performer".

Table 1: Dyads distributed into three levels of performance (low, medium, high) shown by the worst performing individual within each dyad. The benefit of fusion increased as the the difference in baseline accuracy of the paired individuals decreased. ** p < .001

Test	Performance Range	Pearson Correlation
EFCT	Low ($0.667 < \text{AUC} \leq 0.772$)	-0.7930**
EFCT	Medium ($0.772 < \text{AUC} \leq 0.877$)	-0.7979**
EFCT	High ($0.877 < \text{AUC} \leq 0.982$)	-0.6230**
FET	Low ($0.400 < \text{AUC} \leq 0.601$)	-0.5243**
FET	Medium ($0.601 < \text{AUC} \leq 0.800$)	-0.6800**
FET	High ($0.800 < \text{AUC} \leq 1$)	-0.5089**

Table 2: Dyads distributed into three levels of performance (low, medium, high) shown by the best performing individual within each dyad. The benefit of fusion increased as the the difference in baseline accuracy of the paired individuals decreased. ** p < .001

Test	Performance Range	Pearson Correlation
EFCT	Low ($0.678 < \text{AUC} \leq 0.785$)	-0.6485**
EFCT	Medium ($0.785 < \text{AUC} \leq 0.891$)	-0.7859**
EFCT	High ($0.891 < \text{AUC} \leq 0.998$)	-0.7730**
FET	Low ($0.406 < \text{AUC} \leq 0.604$)	-0.6451**
FET	Medium ($0.604 < \text{AUC} \leq 0.802$)	-0.7749**
FET	High ($0.802 < \text{AUC} \leq 1$)	-0.7207**

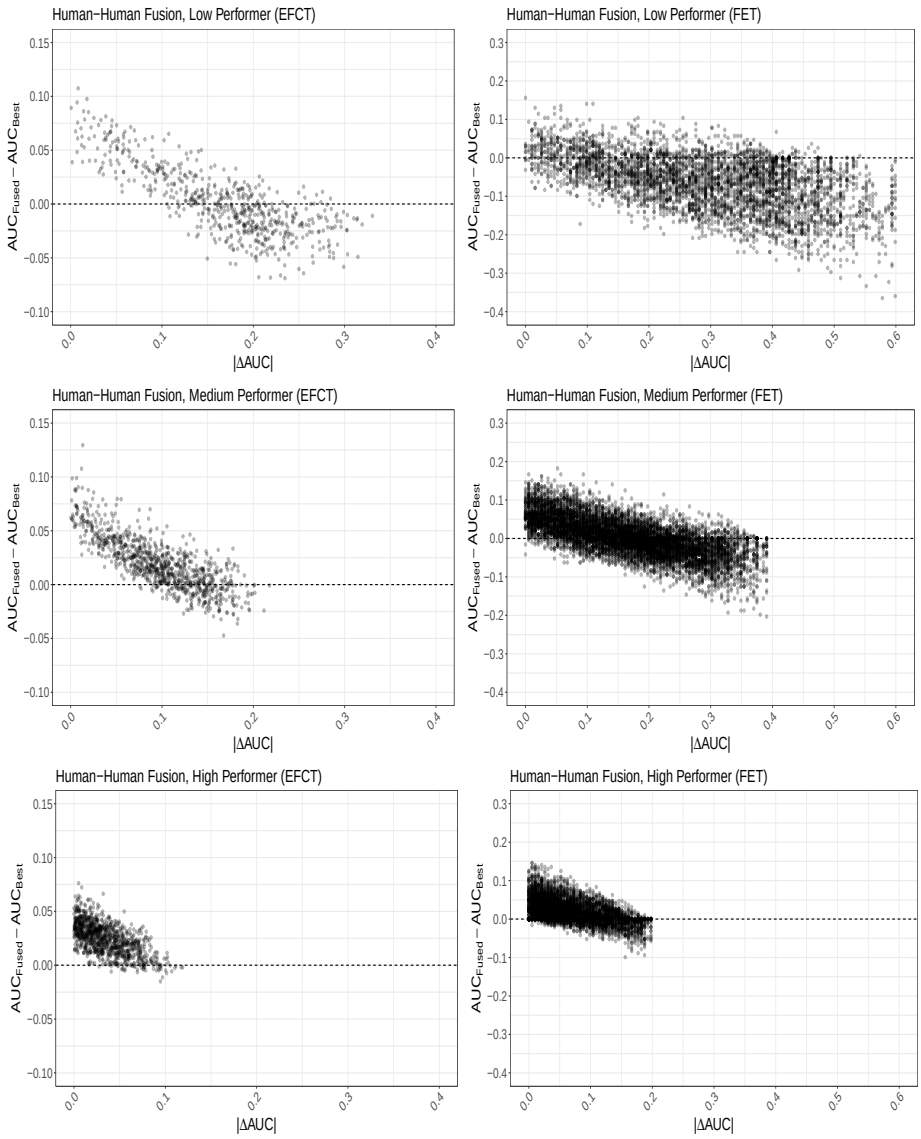

Figure 5: Fusion benefits across different levels of performance shown by the worst performing individual within each dyad. The ranges in performance exhibited by the worst performing member within each dyad were distributed into three levels: low (top), medium (middle), and high (bottom). Data from the EFCT and FET are shown on the left and right, respectively.

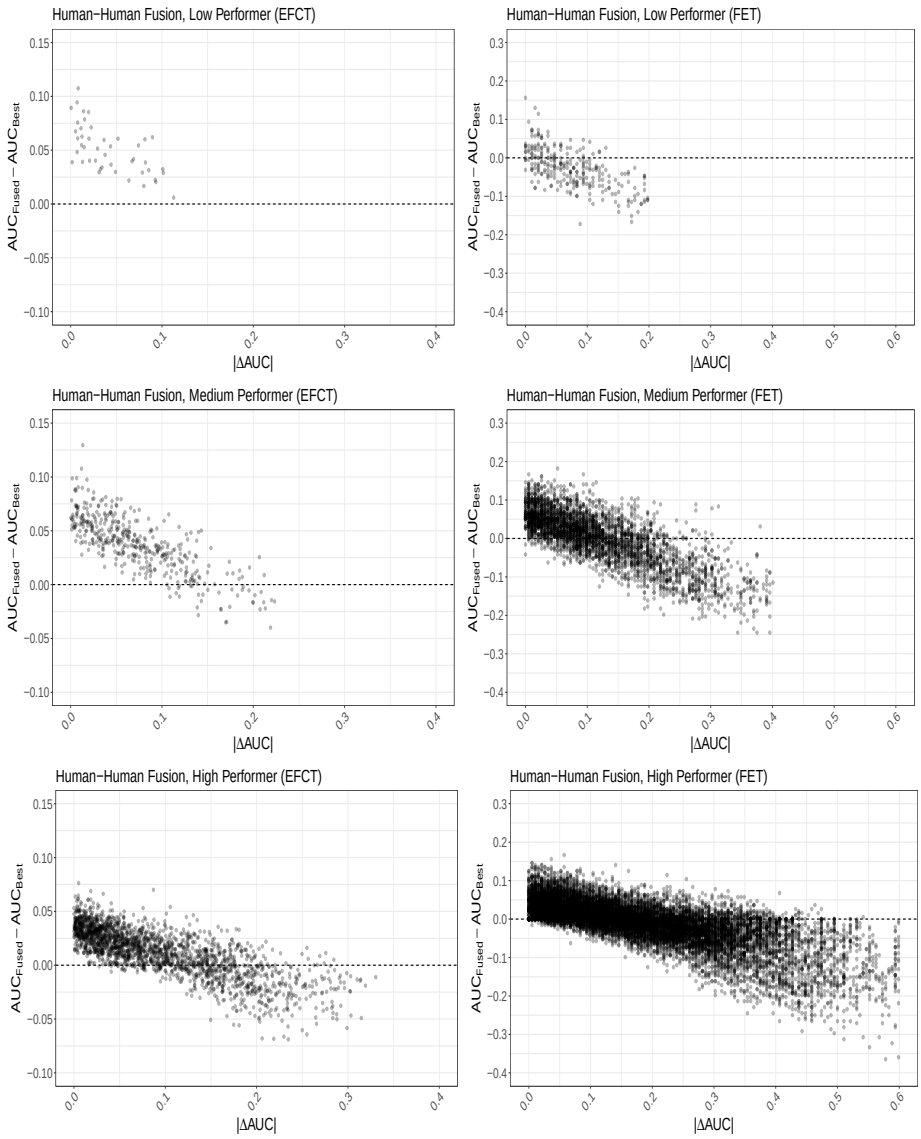

Figure 6: Fusion benefits across different levels of performance shown by the best performing individual within each dyad. The ranges in performance exhibited by the best performing member within each dyad were distributed into three levels: low (top), medium (middle), and high (bottom). Data from the EFCT and FET are shown on the left and right, respectively.