Measuring Error Alignment for Decision-Making Systems

Binxia Xu*, Antonis Bikakis, Daniel F.O. Onah, Andreas Vlachidis, Luke Dickens*

Dept. of Information Studies, University College London, United Kingdom
{b.xu, l.dickens}@ucl.ac.uk

Abstract

Given that AI systems are set to play a pivotal role in future decision-making processes, their trustworthiness and reliability are of critical concern. Due to their scale and complexity, modern AI systems resist direct interpretation, and alternative ways are needed to establish trust in those systems, and determine how well they align with human values. We argue that good measures of the information processing similarities between AI and humans, may be able to achieve these same ends. While *Representational alignment* (RA) approaches measure similarity between the internal states of two systems, the associated data can be expensive and difficult to collect for human systems. In contrast, *Behavioural alignment* (BA) comparisons are cheaper and easier, but questions remain as to their sensitivity and reliability. We propose two new behavioural alignment metrics *misclassification agreement* which measures the similarity between the errors of two systems on the same instances, and *class-level error similarity* which measures the similarity between the error distributions of two systems. We show that our metrics correlate well with RA metrics, and provide complementary information to another BA metric, within a range of domains, and set the scene for a new approach to value alignment.

Code — https://github.com/xubinxia/error_align
Extended version — https://arxiv.org/abs/2409.13919

Introduction

With significant advancements in AI development, the alignment of AI with human values has increasingly drawn attention within the community. Generally, AI alignment focuses on aligning the performance of AI systems towards goals (Zhuang and Hadfield-Menell 2020; Ngo et al. 2022; Sanneman and Shah 2023), preferences (Stray 2020), and social norms (Irving and Askell 2019; Gabriel and Ghazavi 2021) intended by humans. Improved human alignment can help build more reliable and trustworthy AI systems. Considering the potential for AI systems to play a crucial role in future decision-making processes, the trustworthiness and reliability of these systems emerge as critical concerns, particularly in applications such as medical diagnosis and autonomous driving. Studies in cognitive science have demonstrated that a model more closely aligned with the internal mechanism of the human brain can improve the robustness of visual-based decision-making tasks (Dapello et al. 2020). Methods for comparing the internal representations of systems are called representational alignment (RA). RA approaches are often presented as the gold-standard, as similarities between representational structures of systems can provide deep insights about the alignment between the information processing of these systems (Kriegeskorte et al. 2008; Hermann and Lampinen 2020). Nonetheless, RA studies are often limited by the high costs and practical challenges associated with collecting and comparing complex, heterogeneous and difficult-to-access internal representations, e.g. via fMRI in humans. In contrast, the study of how a system behaves can also inform us about both animal and human systems (Staddon and Cerutti 2003; Ghodrati et al. 2014; Rajalingham et al. 2015), as well as computational systems (Geirhos et al. 2019, 2021). However, questions remain about how much behavioural alignment (BA) approaches can tell us about deeper similarities in the internal processing of systems (Hermann et al. 2020; Sucholutsky et al. 2023).

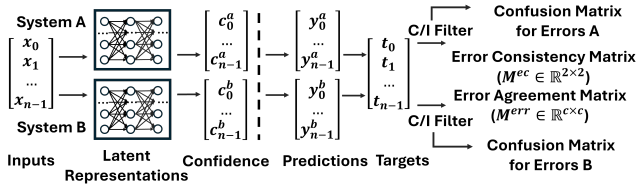

Figure 1: Different levels of representations. From left to right, it enables the comparison of the decision-making process of two systems at the latent representation level, confidence level, instance level and class level.

The separation between RA and BA is not a strict one (Sucholutsky et al. 2023), and there are finer distinctions in the types of observations on a system's information processing pipeline used to evaluate alignment. Figure 1 shows different levels at which observations can be drawn from a typical machine system, with analogous choices for human or animal models. Left of the dotted line are what we might call internal representations and right might be termed behaviours. Here we describe RA approaches as those based

purely on internal representations, and we further divide these into latent representations - the layer activations before the soft-max operations and confidences (soft-max logits). Techniques such as Canonical Correlation Analysis (CCA) (Raghu et al. 2017) and Centered Kernel Alignment (CKA) (Kornblith et al. 2019) have been used to assess the similarity of latent representations, while elsewhere comparisons between model confidences are used (Guo et al. 2017; Papyran et al. 2020). In contrast, we describe BA metrics as those based on observations drawn from right of the dotted line. BA approaches include the error consistency (EC) scores proposed by Geirhos et al. (2020) and the matrices based on object discrimination tasks proposed by Rajalingham et al. (2015). Some studies compare representations of some systems with behaviours of others (Ghodrati et al. 2014; Peterson et al. 2018; Lee et al. 2024).

As a widely used metric for BA, EC measures the degree to which two systems make correct or incorrect predictions simultaneously. However, only measuring when two systems make errors can be problematic in revealing the similarity of decision-making mechanisms, as whether an instance is correctly predicted can highly depend on the uncertainty carried by the data sample. Additionally, imagine a scenario where two systems achieve almost the same accuracy, but one classifies bears as cats, and the other classifies bears as books. Which system is more reliable, or which type of mistake is more acceptable to humans? Therefore, we argue that the alignment should be assessed not only based on when errors occur but also on how errors are made. In this work, we propose two evaluation metrics: Misclassification Agreement (MA) and Class-Level Error Similarity (CLES), to measure how similarly two systems make errors at instance-based and class-based levels - both based on behavioural observations. We argue that people tend to believe two systems have similar decision-making strategies if they both misclassify one instance to the same wrong class, and MA is constructed to be sensitive to this. In constrast, CLES represents a comparison between two systems' error distributions, instead of an instance-by-instance comparison.

We conduct an extensive series of evaluations to determine the value of the new measures we develop. Through the experiments, we show MA captures different information from the previously proposed EC which has a comparable level of access. We also demonstrate that CLES relaxes demands on data access and hence can be applied more widely, even comparing with historical models and data where only the confusion matrices are available. Both of our measures have a strong correlation with more privileged measures, although the correlation of MA is stronger. Additionally, they can also be used as an auxiliary loss for the training of a more human-aligned model. Our evaluations, including with our new measures, shed light on how similar models are in how they make errors and how similar (or rather disimilar) these errors are from other systems. In total, our results advance the understanding of the information processing strategies behind these models' predictions, including differences between synthetically distorted and naturally occurring data. Our measures could facilitate low cost evaluation of human error alignment which can help to explain existing models' errors, and develop new models which better align with human errors. We argue that, as do others, models that have better alignment with humans are inherently more trustworthy (Liu et al. 2023).

The central contributions of this work are two-fold. 1) We propose two evaluation metrics of error patterns to measure the behavioural alignment of two systems: Misclassification Agreements (MA) and Class-Level Error Similarity (CLES). MA quantifies the similarity in how systems make mistakes at the instance level. In contrast, CLES operates at the class level, offering greater flexibility than both MA and EC, particularly when instance-level comparisons are difficult to achieve. 2) We report on comprehensive experiments on four different datasets on vision tasks: one synthetic dataset with a number of subsets, and three naturalistic challenging dataset, including both object recognition tasks on images and human activity recognition tasks on videos, to show the effectiveness and generalisability of those metrics. We show that MA can be a complementary metric for EC, while CLES can be a more flexible proxy for MA. The results also demonstrate that behavioural alignment can reflect the internal representational alignment to a certain degree.

Metrics of Error Alignment

As argued by Geirhos et al. (2020), investigating whether two systems consistently make errors on the same stimuli can help to investigate the similarity of decision-making strategies behind the response. They propose error consistency (EC), to measure behavioural similarity between two classification systems in these terms. More precisely, consider dataset $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^{N}$ comprising inputs in data (stimulus) space $x_n \in \mathcal{X}$, and target labels in finite label space $t_n \in \mathcal{Y}$ of C classes. Let A (resp. B) be a classification system that makes prediction y_n^A (y_n^B) for each input x_n [1]. On $\mathcal{D}$, there are N_c jointly correct instances ($t_n = y_n^A = y_n^B$), and N_e jointly incorrect instances ($y_n^A \neq t_n \neq y_n^B$). The *observed error overlap* is the proportion on which A & B agree, $p_{obs} = (N_e + N_c)/\, N$. This is contrasted using the *error overlap expected by chance*:

$$p_{exp} = p_a p_b + (1 - p_a)(1 - p_b)$$

where p_a (p_b) is the accuracy of A (B). And the EC measure is Cohen's kappa (κ) (Cohen 1960) based on these values:

$$EC(A, B) = \frac{p_{obs} - p_{exp}}{1 - p_{exp}}. \tag{1}$$

We argue that EC captures only one aspect of the behavioural alignment between systems A and B, and propose two complementary metrics. To see this, consider the conceptual representation of data space, $\mathcal{X}$, and associated regions shown in Figure 2. The left image shows data space in which datapoints are deterministically associated with

[1]Most simply, a deterministic predictor g, will predict $y_n^g = y^g(x_n)$, for all x_n. In general, both human and machine systems can give different predictions for each presentation of the same stimulus. In this case, two presentations of the same stimulus are treated as two different datapoints.

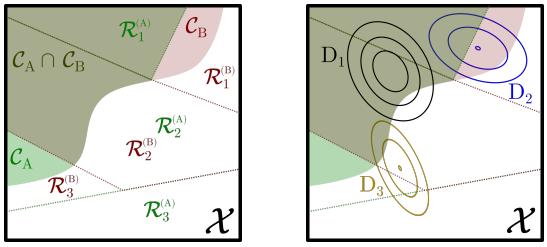

Figure 2: (left) An illustration of dataspace, $\mathcal{X}$, showing decision regions for systems A and B across three classes 1, 2 and 3. Dotted lines indicate decision boundaries for system A (green) and B (red), and decision regions are labelled. The region where system g makes correct classifications, $\mathcal{C}_g$, is shaded appropriately. (right) The same data space, with 3 data distributions, D_i, indicated in black, blue and yellow.

ground-truth classes from $\mathcal{Y}$. For simplicity, we also assume each system g deterministically classifies each datapoint $x \in \mathcal{X}$, with label $y^g(x) \in \mathcal{Y}$. This partitions dataspace into regions $\mathcal{R}_c^{(g)} \subseteq \mathcal{X}$, such that for each system g and class $c \in \mathcal{Y}$ then $y^g(x) = c \Leftrightarrow x \in \mathcal{R}_c^{(g)}$. Similarly, we presuppose that ground truth label for x is determined its position in dataspace, and define for each system g a correct region $\mathcal{C}_g \subseteq \mathcal{X}$, containing all datapoints that would be correctly classified by system g [2]. Figure 2 also shows correct regions for the two systems A & B along with their intersection. Note that, inside the jointly correct region $\mathcal{C}_A \cap \mathcal{C}_B$ (shown in brown) decision boundaries for both systems must necessarily align. Put simply, to jointly correctly classify datapoints they must also agree between themselves. Equally, where one system is correct and another incorrect, e.g. $\mathcal{C}_A \setminus \mathcal{C}_B$ (shown bright green), the systems must necessarily disagree. In the region where both systems are incorrect, $\overline{\mathcal{C}_A \cup \mathcal{C}_B} = \mathcal{X} \setminus (\mathcal{C}_A \cup \mathcal{C}_B)$ (shown in white), there are no such constraints on decision boundaries and they can agree or disagree arbitrarily.

Under these conditions, EC would be calculated solely on the counts of datapoints arising in four distinct regions: both systems correct (brown), A correct B not (green), B correct A not (red), and neither correct (white). Different data distributions for the prediction set give different expected counts within these four regions and hence a different expected value for EC. For example, the right image of Figure 2 shows different distributions as isoprobability contours. Here D_1 and D_3 would have high expected EC between A and B while D_2 would not. This example also illustrates a significant limitation of EC. While distribution D_3 would tend to give high EC scores between A & B, it also gives high probability to data on which A & B disagree. This is because EC treats all data within the dual-misclassification (white) region identically, while the decision boundaries in this region of the two systems can vary substantially.

[2]Note that these simplifications are for the purposes of clarity of illustration, and do not represent limits to the application of EC or our metrics.

Misclassification Agreement

To directly address this insensitivity to differences in the dual-misclassification region, we propose a novel metric called Misclassification Agreement (MA). We first motivate this with an everyday analogy – student performance in exams. In place of systems performing classification tasks, consider students engaged in a multiple choice quiz. When some students answer a question correctly and others incorrectly, we can draw certain inferences about students' learning from when questions are answered correctly and when they are not. This is the essence of the EC metric. Conversely, when a question is answered incorrectly by two or more students, we must consider which incorrect choices each student makes to draw further conclusions about their learning. Moreover, statistical patterns in jointly incorrect choices may expose more about the underlying decision-making of those students. If two students consistently choose the same wrong options, they may share similar misunderstanding about the course content. Equally, this may suggest some ambiguity or common misreading of the question. In a similar way, studying error patterns in classification agents helps us understand the decision-making strategies of those systems. To this end, our novel MA metric offers insight on error patterns by measuring how closely two systems errors align instance-by-instance.

More precisely, with terms A, B, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{D}$, x_n, t_n, y_x^A and y_x^B taking their same meanings as above. The *error dataset* for system g, $\mathcal{D}_g^{\text{err}}$, is those data on which g disagrees with ground truth, i.e. $\mathcal{D}_g^{\text{err}} = \{(x_n, t_n) \in \mathcal{D} : y_x^g \neq t_n\}$, while the joint error dataset between A and B, $\mathcal{D}_{A,B}^{\text{err}}$, is the intersection of the two systems' error datasets, i.e. $\mathcal{D}_{A,B}^{\text{err}} = \mathcal{D}_A^{\text{err}} \cap \mathcal{D}_B^{\text{err}}$. For our example in Figure 2, $\mathcal{D}_{A,B}^{\text{err}}$ corresponds to data arising in the white region ($\overline{\mathcal{C}_A \cup \mathcal{C}_B}$).

We define the *error agreement matrix*, $M_{A,B}^{\text{err}} \in \mathbb{Z}_+^{C \times C}$, as the frequency counts of joint error predictions from systems A and B. More precisely, the (i, j)th cell counts the number of joint error instances x_n where $y_n^A = i$ and $y_n^B = j$, i.e.

$$[M^{\text{err}}]_{ij} = \left|\{(x_n, t_n) \in \mathcal{D}_{A,B}^{\text{err}} : y_n^A = i, y_n^B = j\}\right|$$

The MA between A and B is then the multiclass Cohen's κ (1960) of M^{err}:

$$\text{MA}(A, B) = \kappa(M^{\text{err}}) = \frac{\tilde{p}_o - \tilde{p}_e}{1 - \tilde{p}_e}$$

Here $\tilde{p}_o$ is the observed error-agreement rate between systems, and $\tilde{p}_e$ the probability of chance error-agreement under the null hypothesis that agreement is uncorrelated. Note that these are calculated only over data in $\mathcal{D}_{A,B}^{\text{err}}$. Otherwise, these follow the Cohen's (1960) definitions.

As such, $\tilde{p}_o$ is the proportion of joint errors on which A & B agree, $\tilde{p}_o = N_O^{\text{err}} / N^{\text{err}}$, with $N^{\text{err}} = |\mathcal{D}_{A,B}^{\text{err}}|$ the total number of joint errors and N_O^{err} the number on which A & B agree. Equivalently, $\tilde{p}_o$ is the fraction of counts appearing in the main diagonal of M^{err}. A higher (lower) $\tilde{p}_o$ indicates that two predictors tend to make more (fewer) of the same types of errors. Similarly, probability of chance agreement under the null hypothesis, $\tilde{p}_e$, is what we would expect if the

two predictors predicted independently on error set, $\mathcal{D}^{\text{err}}_{A,B}$:

$$\tilde{p}_e = \sum_{i=1}^{C} \hat{p}_i^{(A)} \cdot \hat{p}_i^{(B)}$$

where $\hat{p}_i^{(g)}$ is the estimated probability that a random element of $\mathcal{D}^{\text{err}}_{A,B}$ is predicted as class $i \in \mathcal{Y}$ by system $g \in \{A, B\}$. For A (resp. B), this is the fraction of counts appearing in the ith row (resp. column) of M^{err}.

Class-Level Error Similarity

Unlike MA, which is based on instance-level comparisons (referred to as trial-by-trial by Geirhos et al. (2020)), our second metric *Class-Level Error Similarity* (CLES) seeks to measure the similarity between predictions of two systems A & B at the class-level, again based on system errors. More precisely, we define, for system $g \in \{A, B\}$, the *error confusion matrix*, F^{err}_g, as the counts of actual (ground-truth) and predicted class on g's error instances, i.e. those in $\mathcal{D}^{\text{err}}_g$. This has (i, j)th element:

$$\left[F^{\text{err}}_g\right]_{ij} = \left|\{(x_n, t_n) \in \mathcal{D}^{\text{err}}_g : t_n = i, y^g_n = j\}\right|$$

Note that by design the diagonal elements $\left[F^{\text{err}}_g\right]_{ii} = 0$.

We then calculate a row-wise Jensen-Shannon divergence (JSD) (Vajda 2009), between two systems, by first converting each row of each matrix to a categorical probability distribution as the expectation of a posterior Dirichlet distribution given prior $\boldsymbol{\alpha} \in \mathbb{R}^C_+$. [3] More precisely, if we collect system g's row i of counts into vector, $\mathbf{f}^g_i$, and define the vector of all 1s as $\mathbf{1}$, this gives estimated error distribution for system g on class i as:

$$\hat{\boldsymbol{\pi}}^g_i = \frac{\mathbf{f}^g_i + \boldsymbol{\alpha}}{\mathbf{1}^T(\mathbf{f}^g_i + \boldsymbol{\alpha})}$$

The *class level error distance* (CLED) between system A and B aggregates these differences as:

$$\text{CLED}(A, B) = \sum_{i=1}^{C} w_i \, \text{JSD}(\hat{\boldsymbol{\pi}}^A_i, \hat{\boldsymbol{\pi}}^B_i) \tag{2}$$

where $w_i = \mathbf{1}^T(\mathbf{f}^A_i + \mathbf{f}^B_i)$.

To make the score comparable to other alignment metrics (and to potentially serve as an auxiliary loss for future human-aligned models) we convert this dissimilarity into a our *class-level error similarity* (CLES) as:

$$\text{CLES}(A, B) = \frac{1}{1 + \text{CLED}(A, B)}. \tag{3}$$

One key aspect of the CLES metric is that the confusion matrices are derived from each system g's error dataset, $\mathcal{D}^{\text{err}}_g$. In terms of Figure 2, CLES is estimating the error distribution of system A's (resp. B's) predictions over the white and red (resp. green) regions. Thus, it combines counts used by both EC and MA, but then measuring similarity between distributions rather than using instance-by-instance agreements. Moreover, this metric is much less sensitive to the degree of difficulty of a given domain compared to a comparison between conventional confusion matrices (Rajalingham, Schmidt, and DiCarlo 2015; Kheradpisheh et al. 2016b,a). To see why, note that if two systems both classify a datapoint correctly, they must necessarily agree on the label (as previously discussed). Hence, a domain in which both systems have a high accuracy will result in two confusion matrices with a high proportion of counts in the main diagonals, which skews similarity comparisons. Another key aspect of this approach is the use symmetric information theoretic measure, the JSD, to evaluate the difference between distributions derived from rows of the confusion matrices. This has the dual advantages of taking account of the non-euclidean geometry of the space in which these predictive distributions sit, and having a meaningful information-theoretic interpretation of the difference between distributions. JSD is preferred over KL-divergence as it is symmetric, and handles zero probabilities more gracefully (Vajda 2009). More details are given in Appendix 1 and 2.

[3] Note that we use Dirichlet Prior with shape parameter $\boldsymbol{\alpha} = 0.5 \cdot \mathbf{1}$ throughout our experiments.

Experiment

This section describes evaluations on our two new BA metrics, alongside pre-existing BA and RA metrics. We aim to investigate what these metrics can tell us about the similarities between pairs of systems both within, and across domains, including human and deep neural network systems. We also wish to evaluate whether, and to what degree, different metrics provide overlapping or complementary information about the similarities between systems. This includes, to our knowledge, the first systematic study on correlations between BA and RA metrics across a range of settings, datasets and system types.

Dataset and Metrics

Our experiments include two groups of classification datasets: synthetic and naturalistic. The synthetic group contains 14 subsets of the `modelvshuman` image dataset (Geirhos et al. 2021), with both machine and human predictions. The naturalistic group comprises three challenging datasets: one image recognition task – `ImageNet-A` (Hendrycks et al. 2021b)), and two video-based Human Activity Recognition (HAR) tasks – `MPII-Cooking` (Rohrbach et al. 2016) and Epic-Kitchen (Damen et al. 2018). For the image datasets, we evaluate a selection of ImageNet1K pre-trained models, and use the complete `modelvshuman` and `ImageNet-A` data for testing. For video datasets, models were trained on the corresponding training set. In all cases, alignment measurements are based on corresponding testing sets. For every pair of systems, across all datasets, we evaluate the three BA metrics already described: EC, MA and CLES, alongside 3 RA metrics: CKA (Kornblith et al. 2019); SOC, the average Jenson-Shannon Divergence (JSD) between the confidences of the two systems; and SOCE, the SOC measure applied to jointly incorrect predictions only. CKA is a SoTA RA metric able to measure multivariate similarity between arbitrary representational spaces. SOC (and SOCE) embodies

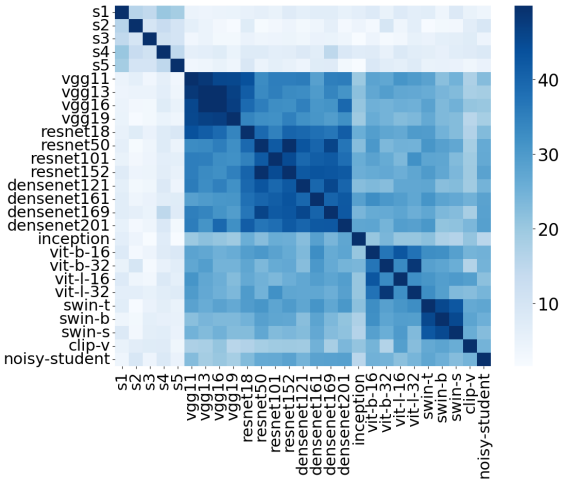

Figure 3: Example heatmap for MA scores on the Stylized subset from `modelvshuman`. Darker cells represent a higher value of similarity.

our own approach to measuring confidence alignment. It is similar in character to the soft cross-entropy loss from Peterson et al. (2019), but SOC is symmetric and more gracefully deals with zero confidences. SOC is also similar to the Hellinger distance based approach from Lee et al. (2024), except SOC better reflects the information geometry of the space of confidences. See more details in Appendix 4 and 5.

Error Patterns between Systems

We evaluate the alignment of each distinct pair of systems by different BA metrics, producing a single score for each model pair on each metric. Figure 3 shows an example heatmap for MA for all pairs of systems on one subset of `modelvshuman`. More heatmaps for different metrics are presented in Appendix 8. In Figure 3, the first five rows/-columns correspond to humans, followed by CNN-based models then transformer-based models. Relative comparisons within one heatmap indicate which pairs of systems are more/less similar according to that metric. Some consistent patterns emerge, for example both human-human and model-model pairs tend to exhibit higher values compared to human-model pairs. This observation aligns with the conclusion drawn by Geirhos et al. (2020), which states that the prediction behaviour between humans and models is less aligned. However, there are differences between metrics and datasets. For example, the MA values for human-human pairs are significantly lower than most model-model pairs within the dataset, whereas EC does not show this same trend, suggesting that EC and MA can measure different aspects of decision-making systems.

EC vs MA. For a more comprehensive investigation of this difference between MA and EC, we systematically compare these alignment scores for all distinct pairs of systems across all subsets in `modelvshuman`. Figure 4 (left) is a scatter plot of these data, where each dot represents a pair of systems on a given subset in `modelvshuman`, such that

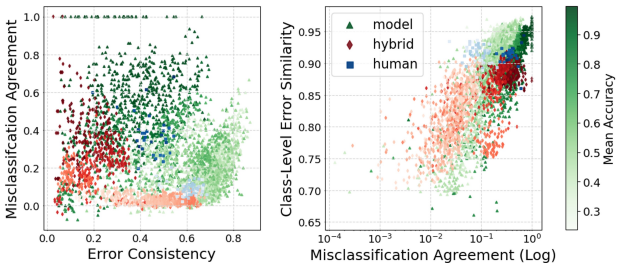

Figure 4: EC vs MA (left) and MA vs CLES (right) on `modelvshuman` data, with model-model, model-human and human-human pairs coloured differently, and shaded according to mean accuracy of the pair.

the x-position is the MA score, and the y-position is the EC score, on that subset. Model-model comparisons are shown in green, human-human in blue, and model-human (hybrid) in red. The colour saturation of these points indicates average accuracy for the two systems on the subset.

Note that, a significant number of points exhibit high EC but low MA, and others have high MA but low EC, substantiating the previous indication that these metrics are complementary. Specifically, pairs with high EC and low MA indicate a high level of agreement on which instances they make errors, but less agreement on which incorrect classes are predicted. Most of the dots in this cluster represent comparisons on highly-corrupted subsets where both systems have low accuracy. We argue that the high EC here, rather than indicating similar decision-making strategies, may stem from a lack of test examples in regions of data space where one system performs better than the other, such as with distribution D_3 from Figure 2 (right). Conversely, the points with medium to high MA but very low EC represent those systems with some agreement on joint errors, but with substantial disagreement on which points are predicted correctly. Most points here correspond to high mean accuracy, and may stem from a lack of test examples in the joint error region, such as with distribution D_1 from Figure 2 (right). It is also worth noting that if we consider only darker dots (those with higher mean accuracy) there is a stronger correlation between the measures, likewise if we restrict ourselves only to lighter dots (those with lower mean accuracy).

This potential sensitivity of EC to low accuracy domains and MA to high accuracy, and the complementarity of measures, advises some caution when interpreting these measures (particularly in isolation). Nonetheless, we can read off some patterns in terms of model-model, human-human and hybrid comparisons here. The emerging picture is that the most highly (behaviourally) aligned system pairs are model-model, while other model-model pairs, as well as human-human pairs, tend to exhibit intermediate levels of alignment, and hybrid pairs show the weakest levels of alignment.

MA vs CLES. Recall that CLES, like MA, measures error prediction similarity between systems, but at the distributional level, rather than instance-by-instance. Figure 4 (right) compares the log(MA) and CLES scores, for all system pairs across all subsets of `modelvshuman`. Each point

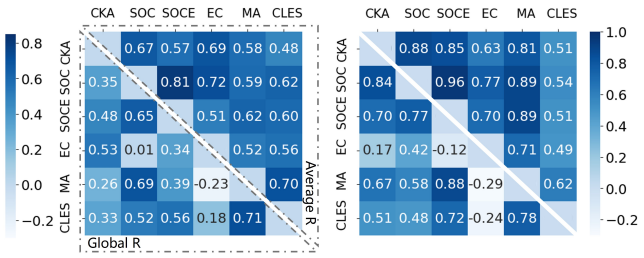

Figure 5: Spearman's r for each pair of metrics for all system-pairs in both synthetic (left) and naturalistic (right) datasets. Global rs measure correlation for all pairs across all datasets; while average rs are the mean in-domain r-value.

represents the two scores for a single system pair coloured as before. Unlike EC vs MA, log(MA) vs CLES exhibits a strong global correlation across all subsets. This relationship is most evident using log(MA) rather than MA, and motivates our use of rank correlation scores in subsequent sections. There is some complementarity here too. For instance, higher mean accuracy points sit more to the right of the group of points (those with relatively higher log(MA) for the same CLES score). This may again be due to MA's sensitivity in low accuracy domains. CLES may be less sensitive as it is drawing data from individual error, rather than joint error, regions. CLES appears to tell a similar story as for EC and MA. The most closely aligned system pairs are model-model. Intermediate alignment is seen with human-human, and some model-model pairs, while hybrid pairs are less aligned. Weaker hybrid scores raise concerns about trustworthiness, and value alignment, of these machine systems.

Correlations across Levels of Alignment

From Figure 4, we observe some evidence of correlation and some complementarity among the BA metrics, and we argue that the complementarity may partly arise due to the influence of domain conditions on different metrics. To investigate this, we calculate in-domain Spearman's r between pairs of BA metrics EC, MA and CLES. Both synthetic and naturalistic datasets consistently show (typically strong) positive correlations between those metrics within the domain (See Appendix 6 for details). Then, to explore these correlations more systematically and extend them to representations as well as behaviours, we consider the correlation between pairs of metrics (both BA and RA) within and across domains, as discussed in the next section.

Global and Average rs. Figure 5 shows the global r (lower triangular) and the average r (upper triangular) between each pair of metrics on the synthetic dataset (left) and the naturalistic dataset (right) - description in caption. Alongside the BA metrics, we consider three RA metrics: CKA, SOC and SOCE. Note that, all metrics align with one another (are rank correlated) moderately or more strongly both within and across domains. We also see consistently lower global r than average r for almost all pairs of metrics on both synthetic and naturalistic data, reinforcing the view

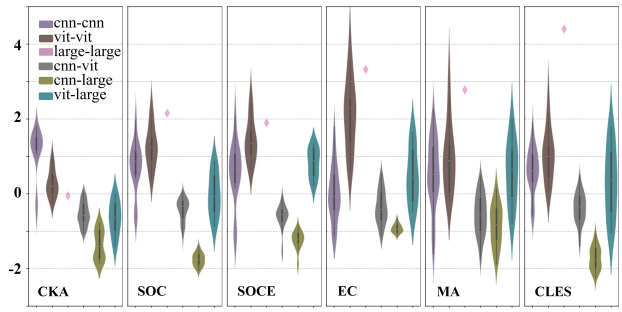

Figure 6: The pair-wise z-scores for different families of models measured by RA and BA metrics for `ImageNet-A`.

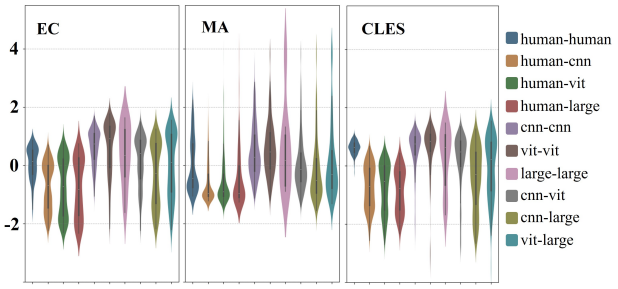

Figure 7: The pair-wise z-scores for different families of systems measured by BA metrics for `modelvshuman` dataset.

that metrics are sensitive to input distributions. In addition, global rs are substantially lower in the synthetic domain. This supports arguments made by Sucholutsky et al. (2023) that naturalistic domains are more likely to express features close to those of the training distribution.

We now address the question: to what extent are these metrics measuring the same thing? Overall, average rs show strong, or very strong, in-domain correlations. This indicates that under a broad set of conditions, high behavioural alignment between two systems provides strong evidence that two machine systems are also closely representationally aligned. Note, however, that these BA vs RA comparisons are based on machine-machine comparisons only.

Alignment across Metrics. Figure 6 and 7 show the pair-wise z-scores for metrics within and across agent families on `ImageNet-A` and `modelvshuman` dataset respectively, where z-scores standardise scores within a metric to facilitate comparisons between metrics. This shows similar patterns of relative similarity within and across agent families across different metrics, although some differences do occur. For `ImageNet-A`, there are three families of models to be compared: CNN, ViT, and larger ViT pretrained on large datasets (Large). Focusing on CKA (the de-facto gold standard), the representations of CNNs are observed to be more similar to each other in the latent space compared to other groups (within and across families), as evidenced by the relatively higher positioning of the CNN-CNN dots. Notably, there are several CNN-CNN dots positioned distinctly below the main group under CKA, SOC, SOCE, and MA.

These dots represent scores between Inception and other CNNs, strongly suggesting that Inception employs substantively different decision-making strategies compared to its CNN counterparts. EC appears to be less effective in capturing these differences between Inception and other CNN models. Moreover, looking at RA metrics, there appears to be an ordering of alignments from CNN-CNN to CNN-ViT to CNN-Large, but in EC and MA, these three groups overlap considerably, suggesting EC and MA are less sensitive to some representational differences. Another interesting point is that although both ViT and Large models have transformer architectures, they differ significantly in the model and training data sizes between groups, and these differences are sufficient to make Large models less aligned with other transformer models. This is in line with other recent findings that scaling matters (Zhai et al. 2022).

For the `modelvshuman` dataset (see Figure 7), humans are included as an agent family to be compared with other model families. Instead of having a plot for each subset, we aggregate all the system pairs across all subsets as one plot. As we don't have human data for RA, we only include BA metrics. Compared to the `ImageNet-A` from Figure 6 - a naturalistic dataset drawn from a single domain, the `modelvshuman` dataset is synthetic and contains a variety of OOD domains. This may explain why the within groups variation dominates, and the between group differences are small. Findings here are less conclusive. This might be because the instances from different subsets of `modelvshuman` vary so much in terms of the features exploited by the models. In addition, a number of the distributions in Figure 7 are bimodal or very long-tailed (suggesting over-dispersion), and may be further evidence of the multi-domain effects discussed previously. However, there are still some conclusions we can draw from these results. For machine-machine comparisons, there is arguably a weak indication that within-group, similarities are higher than between groups, with CNN-Large tending to show the least similarity. For the human comparison, human-human appears (on average) to be more similar than human-machine but less similar than machine-machine. This is more evident for CLES and EC than for MA. This may have something to do with the pure guesswork likely to take place for many of these synthetic images, where, in some cases, there is almost no human discernable class information (see Appendix 7 for some examples). As MA is based purely on jointly incorrect answers, it may be more influenced by these instances than other metrics.

Related Work

Investigating the alignment of representations of different information processing systems is crucial for the understanding of decision-making strategies behind those systems. In RA studies, researchers have focused on measuring the similarity between internal representations of systems (Kriegeskorte et al. 2008; Kornblith et al. 2019; Sucholutsky et al. 2023). Measuring alignments across individuals usually requires the collection of signals from human brain (Kriegeskorte et al. 2008; Nguyen et al. 2022; Sexton and Love 2022). For the exploration of DNN models, hidden activations or confidences are required. For example, Nguyen et al. (2020) and Raghu et al. (2021) conduct comprehensive comparison for the architectures of different models by use of CKA (Kornblith et al. 2019). The alignment between model and humans can also be used to make the prediction of a DNN model more like humans. Peterson et al. (2019) propose a novel approach that leverages human perceptual uncertainty to improve robustness of DNNs by adjusting the confidence. Several complementary works (Peterson et al. 2018; Geirhos et al. 2018; Feather et al. 2019; Kumar et al. 2020; Muttenthaler et al. 2022) have instead used behavioural patterns to reveal the difference between the predictions of neural network models and human results. The alignment of prediction can be done at the instance (trial-by-trial) level (Geirhos et al. 2021), based on class level behaviours, e.g. confusion matrices (Rajalingham et al. 2015; Kheradpisheh et al. 2016b), or at the semantic level (Xu et al. 2024). We argue that our MA metric complements methods from previous instance level approaches as it captures different features of system-system BA. Moreover, unlike previous class-level approaches, our CLES metric is more sensitive to differences as it excludes the influence of correct predictions and is based on the Jensen-Shannon divergence (JSD), which gives a symmetric, parametrisation-independent difference between error distributions.

Another important aspect of any alignment study is the choice of data. Most previous works focus on only one type of dataset, either synthetic datasets (e.g. Out-of-Distribution datasets) (Peng et al. 2019; Hendrycks et al. 2021a; Yang et al. 2022; Lee et al. 2024) or a naturally occurring datasets (Muttenthaler et al. 2022; Karamanlis et al. 2022). As argued by Sucholutsky et al. (2023), the alignment of representations can significantly depend on the selected dataset, underscoring the need for more general studies across different datasets. Additionally, existing research has primarily concentrated on either internal representations or observable behaviours, leaving the relationship between these two aspects under-explored.

Conclusion

In this work, we propose two new metrics for error alignment: MA and CLES, which measure the similarity of errors between classification systems. We evaluate these metrics under a range of conditions and find they correlate well with other BA and RA metrics, and this includes the first systematic study on BA and RA correlations. In particular, our human-model findings correspond with other recent works indicating current image models to be poorly aligned with humans, although more studies are expected to be conducted on the naturalistic datasets. We argue that the latter provides a route to establishing trustworthiness guarantees for systems based on human alignment. However further studies on human-human or human-machine comparisons are needed to determine whether these BA metrics can act as proxies for RA metrics under these conditions. Additionally, those metrics for errors, which can be influenced by accuracy, might not be reliable for the alignment of behaviours in the tasks where systems have achieved nearly perfect performance.

Acknowledgments

We want to thank Dr Xiaoliang Luo, who provided valuable suggestions for this work.

References

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20(1): 37–46.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*.

Dapello, J.; Marques, T.; Schrimpf, M.; Geiger, F.; Cox, D.; and DiCarlo, J. J. 2020. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 13073–13087. Curran Associates, Inc.

Feather, J.; Durango, A.; Gonzalez, R.; and McDermott, J. 2019. Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems*, 32.

Gabriel, I.; and Ghazavi, V. 2021. The challenge of value alignment: From fairer algorithms to AI safety. *arXiv preprint arXiv:2101.06060*.

Geirhos, R.; Meding, K.; and Wichmann, F. A. 2020. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Geirhos, R.; Michaelis, C.; Wichmann, F. A.; Rubisch, P.; Bethge, M.; and Brendel, W. 2019. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *7th Int. Conf. Learn. Represent. ICLR 2019*, (c): 1–22.

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Thieringer, T.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2021. Partial success in closing the gap between human and machine vision. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P. S.; and Vaughan, J. W., eds., *Adv. Neural Inf. Process. Syst.*, volume 34, 23885–23899. Curran Associates, Inc.

Geirhos, R.; Temme, C. R. M.; Rauber, J.; Schütt, H. H.; Bethge, M.; and Wichmann, F. A. 2018. Generalisation in humans and deep neural networks. *Adv. Neural Inf. Process. Syst.*, 31.

Ghodrati, M.; Farzmahdi, A.; Rajaei, K.; Ebrahimpour, R.; and Khaligh-Razavi, S.-M. 2014. Feedforward object-vision models only tolerate small image variations compared to human. *Front. Comput. Neurosci.*, 8.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; and Others. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 15262–15271.

Hermann, K.; Chen, T.; and Kornblith, S. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 33: 19000–19015.

Hermann, K.; and Lampinen, A. 2020. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33: 9995–10006.

Irving, G.; and Askell, A. 2019. AI safety needs social scientists. *Distill*, 4(2): e14.

Karamanlis, D.; Schreyer, H. M.; and Gollisch, T. 2022. Retinal encoding of natural scenes. *Annual Review of Vision Science*, 8: 171–193.

Kheradpisheh, S. R.; Ghodrati, M.; Ganjtabesh, M.; and Masquelier, T. 2016a. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Sci. Rep.*, 6(1): 32672.

Kheradpisheh, S. R.; Ghodrati, M.; Ganjtabesh, M.; and Masquelier, T. 2016b. Humans and Deep Networks Largely Agree on Which Kinds of Variation Make Object Recognition Harder. *Front. Comput. Neurosci.*, 10.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, 3519–3529. PMLR.

Kriegeskorte, N.; Mur, M.; and Bandettini, P. A. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2: 249.

Kumar, S.; Dasgupta, I.; Cohen, J. D.; Daw, N. D.; and Griffiths, T. L. 2020. Meta-learning of structured task distributions in humans and machines. *arXiv preprint arXiv:2010.02317*.

Lee, J.; Kim, S.; Won, S.; Lee, J.; Ghassemi, M.; Thorne, J.; Choi, J.; Kwon, O.-K.; and Choi, E. 2024. VisAlign: Dataset for Measuring the Alignment between AI and Humans in Visual Perception. *Adv. Neural Inf. Process. Syst.*, 36.

Liu, H.; Chaudhary, M.; and Wang, H. 2023. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives. *arXiv preprint arXiv:2307.16851*.

Muttenthaler, L.; Dippel, J.; Linhardt, L.; Vandermeulen, R. A.; and Kornblith, S. 2022. Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*.

Ngo, R.; Chan, L.; and Mindermann, S. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Nguyen, M.; Chang, A.; Micciche, E.; Meshulam, M.; Nastase, S. A.; and Hasson, U. 2022. Teacher–student neural

coupling during teaching and learning. *Social cognitive and affective neuroscience*, 17(4): 367–376.

Nguyen, T.; Raghu, M.; and Kornblith, S. 2020. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*.

Papyan, V.; Han, X. Y.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proc. Natl. Acad. Sci. U. S. A.*, 117(40): 24652–24663.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 1406–1415.

Peterson, J. C.; Abbott, J. T.; and Griffiths, T. L. 2018. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cogn. Sci.*, 42(8): 2648–2669.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 9617–9626.

Raghu, M.; Gilmer, J.; Yosinski, J.; and Sohl-Dickstein, J. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34: 12116–12128.

Rajalingham, R.; Schmidt, K.; and DiCarlo, J. J. 2015. Comparison of Object Recognition Behavior in Human and Monkey. *J. Neurosci.*, 35(35): 12127–12136.

Rohrbach, M.; Rohrbach, A.; Regneri, M.; Amin, S.; Andriluka, M.; Pinkal, M.; and Schiele, B. 2016. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119: 346–373.

Sanneman, L.; and Shah, J. 2023. Transparent Value Alignment. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, 557–560. New York, NY, USA: Association for Computing Machinery. ISBN 9781450399708.

Sexton, N. J.; and Love, B. C. 2022. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science advances*, 8(28): eabm2219.

Staddon, J. E.; and Cerutti, D. T. 2003. Operant conditioning. *Annual review of psychology*, 54(1): 115–144.

Stray, J. 2020. Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, 3(4): 443–463.

Sucholutsky, I.; Muttenthaler, L.; Weller, A.; Peng, A.; Bobu, A.; Kim, B.; Love, B. C.; Grant, E.; Achterberg, J.; Tenenbaum, J. B.; and Others. 2023. Getting aligned on representational alignment. *arXiv Prepr. arXiv2310.13018*.

Vajda, I. 2009. On metric divergences of probability measures. *Kybernetika*, 45(6): 885–900.

Xu, B.; Bikakis, A.; Onah, D.; Vlachidis, A.; and Dickens, L. 2024. Context Helps: Integrating context information with videos in a graph-based HAR framework. In *Proceedings of 18th international conference on neural-symbolic learning and reasoning*. Springer, Cham.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; and Others. 2022. OpenOOD: Benchmarking generalized out-of-distribution detection. *Adv. Neural Inf. Process. Syst.*, 35: 32598–32611.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12104–12113.

Zhuang, S.; and Hadfield-Menell, D. 2020. Consequences of misaligned AI. *Advances in Neural Information Processing Systems*, 33: 15763–15773.