Mapping Trust in AI: Right Tool, Right Task

Connor Baugh, Kyle Camlic, Charles Etheredge, William Marx, Ph.D.,
CAPT Timothy Hill, USN (ret.), Chanler Cantor
Intuitive Research and Technology Corporation (*INTUITIVE*®)
Huntsville, Alabama
connor.baugh@irtc-hq.com; kyle.camlic@irtc-hq.com; chad.etheredge@irtc-hq.com;
william.marx@irtc-hq.com; timothy.hill@irtc-hq.com; chanler.cantor@irtc-hq.com

ABSTRACT

Over the years, the use of Artificial Intelligence (AI) solutions for high-consequence decision-making tasks has become increasingly compelling. However, the application of AI in these settings present legal, moral, and ethical challenges. Due to their highly complex and opaque decision-making, AI agents tend to be viewed as "black boxes." In addition, testing these models is often difficult as their performance on one set of inputs does not necessarily infer their performance on others, making it difficult to predict unexpected behavior. For these reasons, placing full trust in high-consequence AI decision-making has not been feasible.

This paper is a continuation of our research and development on the topic of trust in AI. At I/ITSEC 2023, our team presented a methodology to evaluate AI agent trustworthiness through behavioral modeling. Our research has continued to progress towards a new methodology used to quantify the degree to which AI-enabled systems can be trusted to operate across various scenarios. In this study we utilize a variational autoencoder and gradient mapping techniques to obtain insight into model reasoning and provide objective trustworthiness metrics with intuitive visualizations for human observers. Our proposed method provides distinction between areas of model strength and weakness across varying inputs, as well as detection of out-of-distribution data.

While performing this study, the authors review and compare our approach with current state-of-the-art methodologies in this domain. This paper includes an updated background and literature survey based on new research. Our findings are presented in the context of an experiment to test the methodology used in our approach. It then builds upon this approach to explain how we have applied this methodology to the test and evaluation of piloted and autonomous aircraft operations.

ABOUT THE AUTHORS

Mr. Connor Baugh is a Data Scientist at *INTUITIVE* working on the Internal Research and Development team (IR&D). He is responsible for identifying preexisting data science techniques and developing novel algorithms that can be utilized to solve unique challenges faced by government and industry. He received his MS from the University of West Florida in Data Science and his BS in Economics from Florida State University. His expertise includes predictive modeling (AI/ML), probabilistic inference, and statistical analysis.

Mr. Kyle Camlic is a Data Scientist at *INTUITIVE*. As part of the IR&D team, he is responsible for exploring innovative algorithms and data science techniques to solve complex problems for internal research and development. He received a BS in Applied Mathematics and a minor in Finance from Auburn University. His experience includes applying probability and statistics and developing time-series forecasting models, mathematical modeling, geographical data science, and machine learning.

Mr. Charles Etheredge is a Software Engineer at *INTUITIVE*. As a member of the research and development team, he is tasked with performing and demonstrating research for the company. In this role, he has participated in a wide variety of research, including data visualization, image analysis, audio analysis, cryptography, Artificial Intelligence, and Trust in AI. He has received bachelor's degrees in both, Computer Science and Business Administration from the University of North Alabama. He has expertise in computer vision, video and metadata encoding solutions, as well as web, cloud, and AI applications.

Dr. William Marx is the Senior Vice President and Chief Technology Officer of *INTUITIVE* in Huntsville, Alabama. He is responsible for planning, managing, and executing research and development programs aligned with the technology priorities of the U.S. military and commercial customers. His experience base and technical portfolio include advanced visualization systems, Big Data analytics, Artificial Intelligence and Machine Learning, Knowledge Based Systems, missile system design, multi-disciplinary design optimization, missile guidance and control, analysis of exo-atmospheric kill vehicles, supersonic aircraft design, and ground and space robotic system design. Dr. Marx received his PhD and MS in Aerospace Engineering from the Georgia Institute of Technology and his BS in Aerospace Engineering with a minor in Mathematics from Embry-Riddle Aeronautical University. He was a NASA Langley Graduate Student Researchers Program (GSRP) Fellow.

CAPT Tim Hill, USN, (ret) is the Director of Central Florida Operations at *INTUITIVE*. He is responsible for efficient operation of the Central Florida office and for representing *INTUITIVE* in Central Florida with all customer sets and industry and academic teammates. He is a retired Navy Officer who served across a wide range of operational, staff, and acquisition assignments, culminating in his final tour commanding the Naval Air Warfare Center Training Systems Division (NAWCTSD). He amassed over 3,200 flying hours and 700 carrier arrested landings in more than 32 aircraft types, including operational assignments in the S-3B Viking and F/A-18F Super Hornet. He earned a BS in Systems Engineering from the U.S. Naval Academy, a MS in Systems Engineering from Johns Hopkins University, and a MS in International Relations from Troy University. He is a graduate of the U.S. Naval Test Pilot School and the Air Command and Staff College. He has published articles and papers on topics ranging from tactical aircraft operations to technical studies and acquisition best practices. He is an active member of the Central Florida community through board service and other similar activities.

Ms. Chanler Cantor is an Area Manager at *INTUITIVE*. She is responsible for managing the Internal Research and Technology portfolio and leads multiple projects associated with Artificial Intelligence and Machine Learning, Big Data Analytics, complex visualization, and medical imagery visualization. She and her team develop technology demonstrators and prototypes which can be showcased to customers to promote awareness and early evaluation and adoption of technologies. She has been awarded nine patents related to big data visualization and other emerging technologies. She received her MS in Systems Engineering from Johns Hopkins University and received her BS in Electrical Engineering from The University of Alabama.

Mapping Trust in AI: Right Tool, Right Task

Connor Baugh, Kyle Camlic, Charles Etheredge, William Marx, Ph.D.,
CAPT Timothy Hill, USN (ret.), Chanler Cantor
Intuitive Research and Technology Corporation (*INTUITIVE*®)
Huntsville, Alabama
connor.baugh@irtc-hq.com; kyle.camlic@irtc-hq.com; chad.etheredge@irtc-hq.com;
william.marx@irtc-hq.com; timothy.hill@irtc-hq.com; chanler.cantor@irtc-hq.com

INTRODUCTION

After finding success in multiple domains across government and industry, the use of Artificial Intelligence (AI) solutions for high-consequence decision-making tasks has become increasingly compelling. However, AI models are commonly viewed as “black boxes” due to their highly complex and opaque decision-making, which makes their integration into these scenarios a challenge. Trust in these systems is crucial for the Department of Defense (DoD) because it underpins the responsible adoption and integration of AI technologies into defense operations. The integration of these technologies into DoD operations requires that AI systems be reliable, equitable, governable, and traceable. This is essential for maintaining decision superiority on the battlefield, where speed and accuracy of AI-assisted decisions is imperative.

Evaluating the trustworthiness of AI systems requires rigorous testing and a clear understanding of their capabilities and limitations across a wide array of scenarios and dynamic environments. However, these evaluations are primarily concerned with prediction accuracy and fail to provide stakeholders with critical information pertaining to model reasoning that may indicate the presence of fundamental flaws or vulnerabilities in a system. With the ability to comprehensively evaluate the trustworthiness of their AI systems, the DoD would be able to better leverage these technologies to support warfighters on the battlefield.

In this paper we propose a novel trustworthiness metric and visualization technique that allow DoD stakeholders to intuitively and comprehensively assess the trustworthiness of mission critical AI systems across various scenarios. Specifically, we utilize a variational autoencoder (VAE) in addition to gradient-based stability and uncertainty quantification techniques to obtain insight into model reasoning and provide a visualization of AI trustworthiness with metrics that directly correspond to the stakeholders’ risk tolerance levels. Using a flight path classification example simulating the test and evaluation of piloted and autonomous aircraft operations, we show that our proposed method is capable of distinguishing between areas of model strength and weakness as well as detecting out-of-distribution (OOD) and outlier data, providing stakeholders with an understanding of where their AI model is the right tool for the task at hand.

PRIOR RESEARCH

Trust is essential for the successful development, deployment, and adoption of any high-consequence decision-making model. However, a singular, widely accepted formulation and mathematical foundation of model trust is lacking in the literature. In this work, we formulate AI trustworthiness as a function of algorithmic stability and uncertainty quantification, taking inspiration from prior contributions in the fields.

Algorithmic Stability

First introduced by Kearns and Ron (1999) as a consideration in bound estimation for generalization error, algorithmic stability has become an important concept in the field of explainable AI for analyzing risk in model decision-making. In their early work on algorithmic stability, Bousquet and Elisseeff (2001) defined a stable model as one that is robust to small changes in the training data. Later, Hardt et al. (2016) expanded upon Bousquet and Elisseeff’s theory of algorithmic stability by showing that neural networks trained using stochastic gradient descent (SGD) for a small number of iterations have vanishing generalization error and that common tactics such as dropout and L2 regularization promote algorithmic stability and thus better generalization.

Borrowing from the notion of algorithmic stability, Fel et al. (2021) proposed two novel metrics to evaluate the trustworthiness of explanations of an AI's decision-making. They argued that while several methods have been proposed to explain how deep neural networks reach their decisions, comparatively little effort has been made to ensure that the explanations produced by these methods are objectively trustworthy. Of the two metrics proposed, Mean Generalizability (MeGe) provided the most definitive interpretation of trustworthy explanations as those which focus on the same evidence for all correctly classified images belonging to that class.

Uncertainty Quantification

Bhatt et al. (2021) argued that quantifying and communicating model prediction uncertainty is crucial for algorithmic transparency. The authors claimed that explanations of AI decision-making alone may not be enough for stakeholders to fully gauge the trustworthiness of model predictions, and that incorporating information about where the model may be wrong or lacks sufficient training to solve a given task provides additional context that is critical for evaluating AI trustworthiness.

In the classification setting, the softmax confidence scores produced by the model appear to provide a convenient estimate of model prediction uncertainty. However, these scores tend to be overconfident and are prone to manipulation by adversarial data, making them particularly unreliable for evaluating high-consequence AI decision-making (Nguyen et al., 2015). For a more reliable estimate, Lee and AlRegib (2020) proposed to instead utilize model gradients for epistemic uncertainty quantification and demonstrated the effectiveness of their method in detecting OOD and corrupted data. Similarly, Malmström et al. (2020) derived a method for estimating the asymptotic prediction error variance for feedforward neural networks which utilizes propagation of uncertainty to provide an estimate of overall model uncertainty, incorporating both aleatoric and epistemic uncertainties. Notably, it can be shown that the authors' derivation can be generalized to any feedforward network trained via least squares minimization or log-likelihood maximization, which comprise most cost functions used in practice.

OUR RESEARCH

Prior Contributions

Our team has focused on continued contribution to industry research on trust in AI since 2021 with prior concentrations on behavioral cloning and analysis of black box AI models. In 2021, our team began exploring the possibility of cloning human decision-making for a simple board game into an AI model. During this effort, we determined that we did not have enough human data to sufficiently replicate the human behavior, but after expanding this effort to include AI models, it was shown that an observer agent can learn to effectively predict the behavior of an observed model when given enough observations to sufficiently cover the problem space (Etheredge et al., 2022). This sparked an interest within the team to determine if the decisions of our cloned models could be trusted, leading to our 2023 publication, where a system to engender trust in an AI model by identifying model behaviors observed during specific conditions and scenarios in autonomous vehicles was explored (Russell et al., 2023).

In this work, we take inspiration from prior research on algorithmic stability and uncertainty quantification to derive a novel metric for assessing the trustworthiness of DoD model predictions. Specifically, we adapt the methods proposed by Fel et al. (2021) and Malmström et al. (2020) to create a unified statistic which accounts for both stability and uncertainty. The resulting metric is then incorporated with our proposed visualization technique to provide an all-encompassing view of model trustworthiness across varying scenarios, allowing for confident and effective decision-making by DoD stakeholders.

Latent Space Generation

Effectively mapping trust in AI requires us to encode our data into a lower dimensional representation, commonly known as an embedding or latent space. This is generally done using dimensionality reduction techniques such as principal component analysis (PCA), uniform manifold approximation and projection (UMAP), and t-distributed stochastic neighbor embedding (t-SNE). However, not all dimensionality reduction techniques are equally viable for our purposes, which require that the latent space preserve the global structure of the high dimensional data.

While methods like PCA and UMAP meet this requirement, we instead choose to use a VAE due to some desirable properties described in Battey et al. (2021). In their paper, the authors demonstrated that VAEs are not only superior to other methods in preserving the global structure of the data, but that their loss functions also incentivize meaningful distances between samples, allowing them to achieve sufficient local clustering. These properties are particularly useful in the context of mapping trust in AI as they allow for the smooth traversal between different regions of the data and enable us to easily identify edge cases where we may encounter transitions between trustworthy and untrustworthy model behavior. The resulting latent space thus provides us with a "model worldview" that comprises the inherent structure or organization of the data being input into the black box AI model.

Two-Dimensional Warping

Here we motivate the adaptation of the approach used by Fel et al. (2021) to measure the generalization of model predictions. Like the prior method, we derive the Spearman rank correlation between sets of explanations. However, we deviate from the authors' proposed use of ensemble methods, i.e. k-fold cross validation, to handle image misalignment problems.

In their work, Fort et al. (2020) demonstrated that the effectiveness of deep ensemble methods is, at least in part, due to the individual networks' abilities to converge toward different local minima in the loss landscape. Therefore, the direct comparison of explanations generated from ensemble networks may not be valid as they are not guaranteed to come from models sharing similar generalization properties. It is also unclear how metrics derived in this manner would explain the generalization properties of a specific black box model in question.

Instead, we utilize the two-dimensional warping algorithm introduced by Uchida and Sakoe (1999) as an alternative method for handling image misalignment. By warping sample images to their nearest neighbors in the latent space, we can map their associated explanations to those of their neighbors, allowing for a direct pixel-to-pixel comparison. Notably, addressing misalignment in this way does not require retraining, and the resulting correlation coefficients provide a clear interpretation of model stability.

Mathematically Defining Trust

In this section we derive our proposed trustworthiness metric as well as a computationally efficient approximation. By showing that the prior contributions of Fel et al. (2021) and Malmström et al. (2020) can be reformulated to have useful statistical properties, we create a unified statistic interpreted as the signal-to-noise ratio measuring the degree of algorithmic stability relative to model prediction uncertainty for a given input. The resulting metric provides the level of confidence DoD stakeholders can have in the trustworthiness of a model's prediction, e.g. a value of 0.95 would indicate that we can be 95% confident that a prediction is trustworthy.

Reformulating Uncertainty

Here we show that the model prediction uncertainty estimate derived by Malmström et al. (2020) can be reformulated as a linear combination of chi-squared distributed sample variances under mild assumptions. By reformulating the uncertainty estimate in this way, we can then approximate its sampling distribution which will provide us with useful properties for deriving our proposed trustworthiness metric. For simplicity of notation, we assume that uncertainty estimates are derived for individual samples.

Consider the following model

$$\hat{y}_n = f(x_n, \hat{\theta}), \tag{1}$$

where $f(x_n, \hat{\theta})$ represents our trained neural network and $\hat{y}_n$ is the network's prediction given input x_n and model parameters $\hat{\theta}$. The approximated covariance matrix of each parameter vector $\hat{\theta}^l$ as defined by the authors is given as

$$Cov(\hat{\theta}^l)_n = \frac{\partial L}{\partial \hat{y}_n}^T \frac{\partial L}{\partial \hat{y}_n} \left[\frac{\partial \hat{y}_n}{\partial \hat{\theta}^l} \frac{\partial \hat{y}_n}{\partial \hat{\theta}^l}^T \right]^+, \tag{2}$$

where $\frac{\partial L}{\partial \hat{y}_n}$ is the Jacobian of the loss function with respect to the prediction, $\frac{\partial \hat{y}_n}{\partial \hat{\theta}^l}$ is the Jacobian of the prediction with respect to the parameter vector, and $+$ denotes the Moore-Penrose inverse. If we assume that the parameters follow a multivariate normal posterior, then the sample variances along the diagonal of the approximated covariance matrix are known to be chi-squared distributed[1] such that $\sigma_i^2 \sim \chi_1^2$.

Because the summation of the sample variances requires independence to remain chi-squared, the covariance terms cannot be directly incorporated. Instead, we apply a unitary transformation to diagonalize the approximated covariance matrix such that

$$Cov_{Diag}(\hat{\theta}^l)_n = \begin{bmatrix} \tau_0\sigma_0^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tau_i\sigma_i^2 \end{bmatrix}, \quad \tau_i = \begin{cases} \lambda_i, & \lambda_i \geq 0 \\ 0, & \lambda_i < 0 \end{cases}, \quad \Lambda = [\lambda_0 \cdots \lambda_i], \tag{3}$$

where Λ is the eigenvalue vector of the covariance matrix in (2) and τ_i is defined by Higham (1988) as the case for deriving the eigenvalues of the nearest positive semi-definite (PSD) covariance matrix in the Frobenius norm. Here, τ_i is necessary to ensure that the sample variances in (3) remain valid, since approximated covariance matrices are not guaranteed to be PSD and therefore may contain negative eigenvalues. In this way, the diagonalized covariance matrix can be seen as the unitary transformation of the nearest PSD matrix to (2) in the Frobenius norm.

By replacing the covariance matrix in (2) with (3), the resulting model prediction uncertainty estimate

$$Var(\hat{y}_n) = \sum_l \frac{\partial \hat{y}_n}{\partial \hat{\theta}^l}^T Cov_{Diag}(\hat{\theta}^l)_n \frac{\partial \hat{y}_n}{\partial \hat{\theta}^l} \tag{4}$$

can thus be expressed as a linear combination of chi-squared distributed sample variances

$$Var(\hat{y}_n) = \sum_l \sum_i \left[\frac{\partial \hat{y}_n}{\partial \hat{\theta}_i^l}^2 \tau_i \right] \sigma_i^2 . \tag{5}$$

Deriving the Trustworthiness Metric

Given the reformulation of the model prediction uncertainty estimate in (5), its sampling distribution can be safely approximated by another chi-squared distribution with the effective degrees of freedom given by the Welch-Satterthwaite equation[2] such that $Var(\hat{y}_n) \sim \chi_v^2$ and

$$v = \frac{Var(\hat{y}_n)^2}{\sum_l \sum_i \left(\left[\frac{\partial \hat{y}_n}{\partial \hat{\theta}_i^l}^2 \tau_i \right] \sigma_i^2 \right)^2}. \tag{6}$$

To incorporate algorithmic stability, we apply Fisher's transformation[3] to our Spearman correlation coefficient ρ_n such that

$$z_n = arctanh(\rho_n) \sim N(0,1). \tag{7}$$

Given (5), (6) and (7), we derive our unified statistic

$$t_{\hat{y}_n} = \frac{z_n}{\sqrt{\frac{Var(\hat{y}_n)}{v}}} \sim t_v, \tag{8}$$

[1] The sample variances are unbiased for the stated model; hence Bessel's correction is unnecessary. However, this is not generally the case if we wish to account for model misspecification.

[2] A study conducted by A. Feiveson and F. Delaney (1968) for the National Aeronautics and Space Administration (NASA) indicates that this approximation can be safely used for equal sample sizes as described.

[3] ρ_n is a population statistic for the respective explanations, therefore its transform is standard normally distributed.

which can be interpreted as a signal-to-noise ratio measuring the degree of algorithmic stability relative to model prediction uncertainty given some input x_n. Since our statistic comes from a known sampling distribution, we can calculate its p-value corresponding to a two-tailed test

$$\Phi_{t_v}(t_{\hat{y}_n}) = 2Pr(t > |t_{\hat{y}_n}|) = 2\left(1 - F_{t_v}(t_{\hat{y}_n})\right), \tag{9}$$

where F_{t_v} is the cumulative density function of the Student's t-distribution with v effective degrees of freedom. The resulting p-value $\Phi_{t_v}(t_{\hat{y}_n})$ provides the probability of committing a type 1 error and can therefore be interpreted as the probability of incorrectly placing trust in the neural network's prediction $\hat{y}_n$.

Given (9), we finally express our proposed trustworthiness metric as

$$Trust = 1 - \Phi_{t_v}(t_{\hat{y}_n}), \tag{10}$$

which provides the confidence in the trustworthiness of the network's prediction $\hat{y}_n$.

Mean-Field Approximation

Here we derive a more computationally efficient approximation to our trustworthiness metric, allowing us to avoid eigendecomposition and matrix inversions. By assuming that the model parameters $\hat{\theta}$ in (1) follow a factorized Gaussian posterior such that each parameter is independent of the others, we can reformulate (2) as

$$Cov_{MF}(\hat{\theta}^l)_n = \frac{\partial L}{\partial \hat{y}_n}^T \frac{\partial L}{\partial \hat{y}_n} Diag\left(\frac{\partial \hat{y}_n}{\partial \hat{\theta}^l}\frac{\partial \hat{y}_n}{\partial \hat{\theta}^l}^T\right)^{-1}, \tag{11}$$

where $Diag\left(\frac{\partial \hat{y}_n}{\partial \hat{\theta}^l}\frac{\partial \hat{y}_n}{\partial \hat{\theta}^l}^T\right)^{-1}$ is simply the reciprocated elements of the diagonalized outer product. Since (11) is already diagonalized, we can avoid eigendecomposition and use the mean-field approximated covariance matrix in place of (3). Consequently, (5) and (6) remain the same with the omission of τ_i.

APPLICATION

In this section we outline our general approach for mapping trust in AI as shown in Figure 1 and apply it to a flight path binary classification problem. Using our VAE and novel trustworthiness metric as defined in the prior section, we illustrate how this approach can be used to produce a visualization of AI trustworthiness across various flight paths. Following our application, we then provide a comprehensive analysis of the resulting trustworthiness map and present additional findings from our experiments that demonstrate the viability of our proposed method.

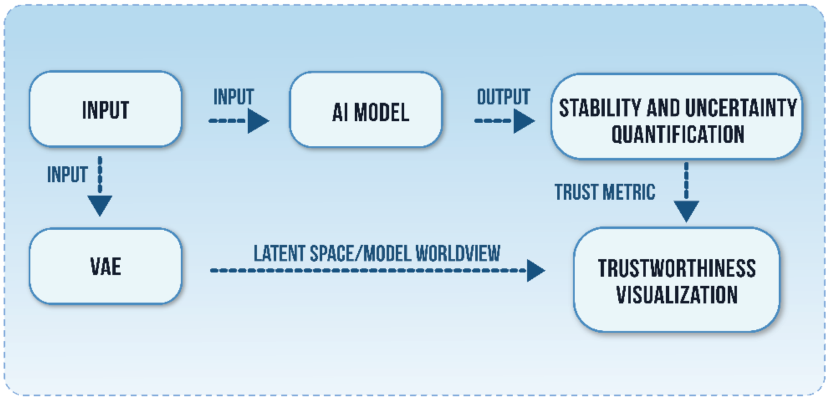
INPUT
INPUT
AI MODEL
OUTPUT
STABILITY AND UNCERTAINTY QUANTIFICATION
INPUT
TRUST METRIC
VAE
LATENT SPACE/MODEL WORLDVIEW
TRUSTWORTHINESS VISUALIZATION

Figure 1. Approach for Mapping Trust in AI

Flight Path Dataset

For our experiments, we use a commercial flight path dataset for a binary classification task to simulate the test and evaluation of piloted and autonomous aircraft operations conducted by the DoD. An example commercial flight path image is displayed in Figure 2.

To generate our dataset, we collected 480 normal flight paths of commercial airliners travelling between Atlanta and Orlando from FlightAware.com, a publicly available flight tracking site. We then synthesized 481 abnormal flight paths by permutating a set of manually generated paths, each deviating from the standard routes by varying degrees. Each of the flights were labeled as either normal or abnormal for training. Lastly, the data was preprocessed for use as model inputs as shown in Figure 3.

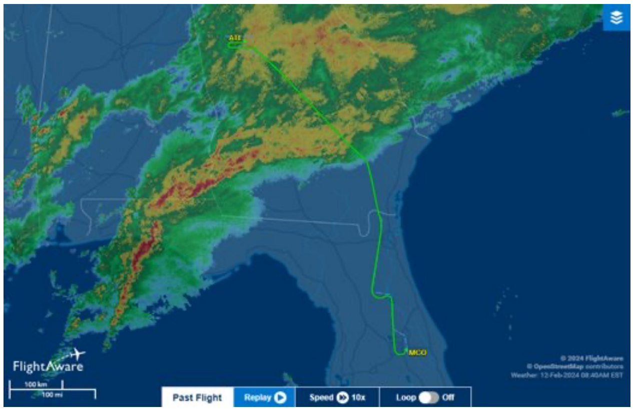

Figure 2. Image of flight path from Atlanta, GA to Orlando, FL from FlightAware.com

Encoding Our Data

As referenced in the prior section, we choose to utilize a VAE to encode our dataset due to its desirable properties for mapping trust in AI. A scatterplot of the resulting latent space as well as a visualization of flight paths as organized in the space are presented in Figure 4.

From the scatterplot it is immediately clear that abnormal flights exhibit a far greater degree of variability than their normal counterparts, as indicated by the differences in the two classes' densities. This aligns with our prior knowledge of the dataset, where normal flights tend to deviate very little from the standard routes. Additionally, the region in the center of the latent space where the two classes are tightly bordered suggests that AI models trained on this dataset will likely suffer from instability or higher generalization error when attempting to classify these edge cases. Our results in the following section confirm these observations.

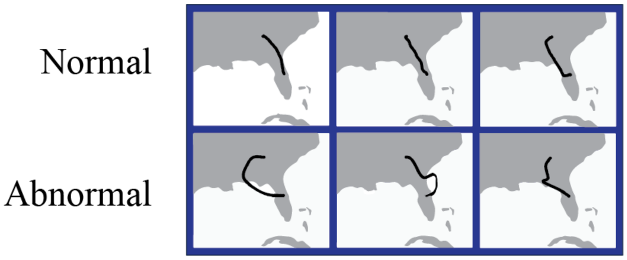

Figure 3. Example Normal and Abnormal Flights

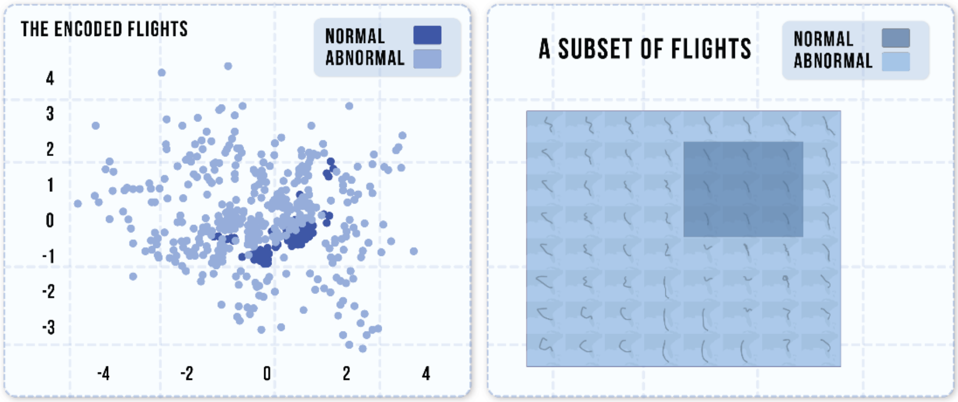

Figure 4. Scatterplot (left) and visualization (right) of flight paths in the VAE latent space

Visualizing Trust in a Black Box Model

For the purposes of our experiments, we train a simple convolutional neural network (CNN) to classify the flight paths in our dataset. Notably, we do not use any regularization techniques to aid training as that would have an impact on algorithmic stability and thus bias our results. Once training is complete, we evaluate the model on each of the flight paths and calculate their associated explanations. Here we choose to utilize Gradient-weighted Class Activation Mapping (Grad-Cam) to calculate model explanations due to its superior performance relative to competing methods as outlined by Selvaraju et al. (2017) and Fel et al. (2021).

After warping each image to their nearest neighbors in the latent space, we then map their associated explanations to align with their neighbors and derive the Spearman rank correlation coefficient to calculate our measure of algorithmic stability as defined in equation 7. Next, we derive the prediction uncertainty estimate and effective degrees of freedom for each image as defined in equations 5 and 6. Finally, given the components outlined above, we calculate the trustworthiness associated with each image as defined in equation 10 and overlay the resulting metrics onto our visualization of the flight paths in the latent space to produce an all-encompassing view of model trustworthiness. The resulting trustworthiness map is shown in Figure 5 in the following section.

RESULTS

Evaluating Trust in Flight Path Classifications

Our results demonstrate strong performance using our proposed method to evaluate the trained binary classification model, as illustrated in Figure 5. In each plot, brighter flight paths indicate areas where the model outputs untrustworthy predictions, is uncertain in its predictions, or exhibits unstable decision-making, respectively. Additionally, red flight paths indicate misclassifications by the model. The values for the stability and uncertainty plots correspond to the numerator and denominator of equation 8, respectively.

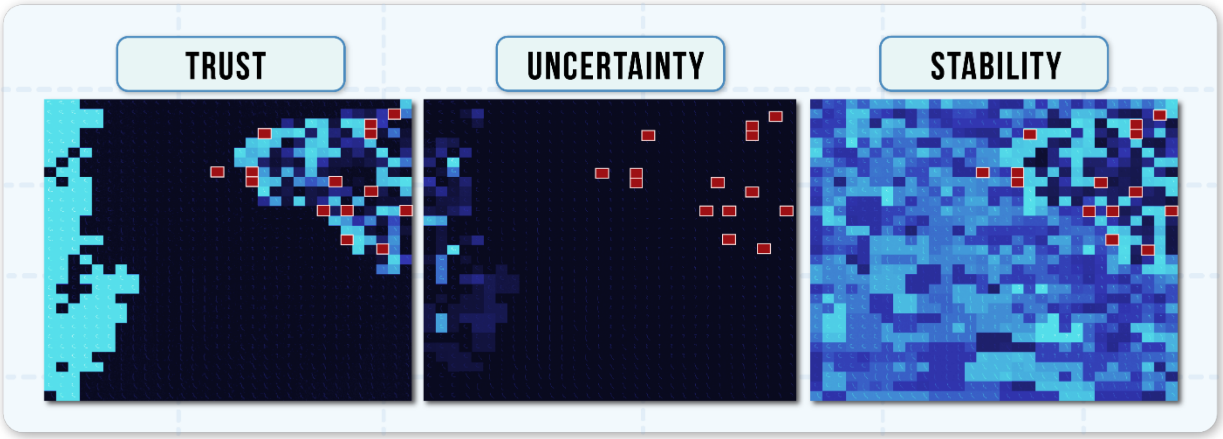

Figure 5. Model Trustworthiness (left), Uncertainty (middle), and Stability (right) across flight path dataset

As shown by the trustworthiness map above, our method was able to clearly delineate between areas of model strength and weakness, identifying two primary regions of low trustworthiness. Here, most abnormal flight classifications are deemed entirely trustworthy, while classifications for edge cases and outliers are found to be less trustworthy to varying degrees. For the purposes of our evaluation, this map provides a high-level breakdown of the scenarios in which the model can and cannot be trusted; however, on its own it does not explain why certain classifications are lacking trustworthiness.

By analyzing the uncertainty map in Figure 5, we can see that the lack of trustworthiness in model predictions corresponding to the leftmost region of the latent space is likely due to model uncertainty in those samples. Here it seems the model was uncertain about flights veering west into Alabama and Mississippi, which deviate significantly from most flight paths in the dataset. This uncertainty indicates a potential need for additional training samples of flights travelling over these states.

Lastly, the stability map displayed in Figure 5 confirms our observations from the previous section as shown by increased instability surrounding the region containing edge cases between the two classes. Interestingly, although the model only exhibits moderate stability across most examples in the latent space, it exhibits far greater stability for normal flight classifications. The general lack of stability exhibited by the model indicates that it likely would have benefitted from regularization during training to increase its generalization power.

Detecting OOD and Outlier Data

Here we demonstrate the effectiveness of our proposed trustworthiness metric in detecting OOD and outlier data. Corresponding to prior research findings on model reliability and generalization, we expect an objective metric to return high trustworthiness around samples common to the dataset and to significantly decrease around outliers and edge cases. Figure 6 presents an examination of trust derived from model flight path classifications in the detection of outliers in the dataset. The box plot visualizes the distribution of the flight path data, where the Euclidean distances are measured from the centroid of the respective flight classes in the latent space. Outliers in the data are denoted as diamonds in the plot. As expected, we find that trustworthiness metrics for smaller Euclidean distances are considerably higher than those for larger distances.

To demonstrate our metric's ability to detect OOD data, we utilize the well-known MNIST and EMNIST datasets which are comprised of images of handwritten digits (0 – 9) and letters, respectively. We train a new CNN to classify the MNIST digits, continuing to avoid regularization techniques, and then evaluate the model on samples from both MNIST and EMNIST datasets for comparison. Here, samples from EMNIST are considered OOD since the MNIST classifier is not trained on the dataset. Due to the inherent uncertainty surrounding never-before-seen samples, we expect there to be little to no overlap between the trustworthiness distributions derived from the two datasets.

As shown in Figure 7, the two distributions diverge toward opposite ends of the spectrum, sharing little overlap as expected. Notably, the overlap between the two distributions corresponding to highly trustworthy predictions can be attributed to the resemblance of some letters in EMNIST to digits in MNIST, e.g. "B" and "8" share similar visual characteristics. Regardless, however, these results signify that our trustworthiness metric is effective in distinguishing between OOD and within-distribution data.

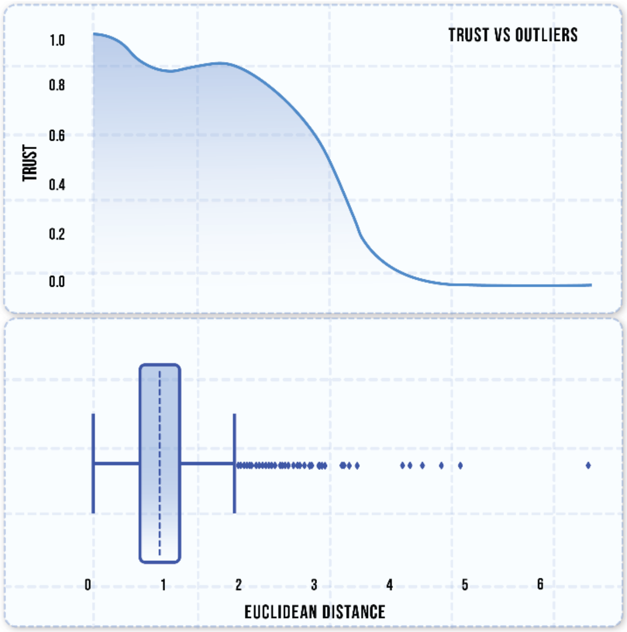

Figure 6. Trust vs. Outliers

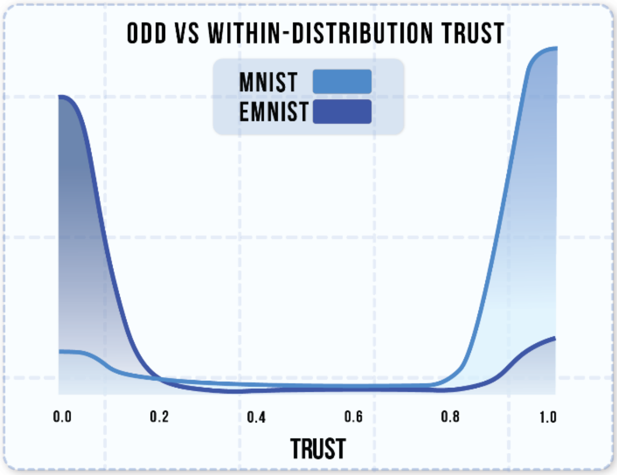

Figure 7. OOD vs. Within-Distribution Trust

CONCLUSIONS AND FUTURE WORK

Our research highlights the importance of reliable AI in defense applications. We've shown that our novel use of VAEs and gradient-based stability and uncertainty quantification techniques offers a robust method for evaluating AI model trustworthiness and pinpointing strengths and weaknesses. As government and industry continue to evolve and increasingly complex models become more prevalent, the ability to trust these models will be crucial to operational effectiveness and decision-making reliability. Methods used to determine trust in AI need to be complex enough to capture information on how a model is reaching its decision, not just that it is correct, and be simple enough for anyone, regardless of background, to understand. This is key to building confidence in these systems and is essential for the widespread adoption and responsible deployment of AI technologies across the DoD.

The method proposed in this paper focuses on the evaluation of a CNN trained on a commercial flight path binary classification problem simulating the test and evaluation of piloted and autonomous aircraft operations. While this method provides a powerful framework for AI trust assessment, it primarily focuses on image data and would require adaptation for other data types. Therefore, extensions of these techniques to AI models intended for non-visual data and further refinements to improve efficiency and scalability may be explored in future work. Additional future work expanding evaluation to real or more realistic data simulating DoD aircraft operations would also be valuable.

The approach postulated in this paper is a novel method for quantifying and visualizing AI model trustworthiness, offering a comprehensive tool that combines latent space mapping and gradient-based prediction trustworthiness assessment to provide actionable insights into model reliability. Rationale behind AI decision-making is vital in order to effectively evaluate AI systems and forecast whether errors are likely to occur across various scenarios. Building and maintaining trust in AI technologies is essential for the DoD to leverage its full potential while enabling personnel to rely on AI-driven recommendations in critical situations. We strongly encourage researchers to expand upon our framework to further improve AI transparency and we implore policymakers to take these advancements into account when setting guidelines for use of AI in high-stakes scenarios.

REFERENCES

Battey, C. J., Coffing, G. C., & Kern, A. D. (2021). Visualizing population structure with variational autoencoders. *G3*, *11*(1). https://doi.org/10.1093/g3journal/jkaa036

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462571

Bousquet, O. and Elisseeff, A. (2001). Algorithmic stability and generalization performance. *Advances in Neural Information Processing Systems*, pp. 196–202. https://proceedings.neurips.cc/paper_files/paper/2000/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

Etheredge, C., Russell, K., Marx, W., Hill, T. & Drown, D. (2022). The use of AI/ML to replicate threat behavior for nonlinear simulation. *I/ITSEC 2022*. https://s3.amazonaws.com/amz.xcdsystem.com/44ECEE4F-033C-295C-BAE73278B7F9CA1D_abstract_File16562/PaperUpload_22230_0630041243.pdf

Feiveson, A., & Delaney, F. (1968). The distribution and properties of a weighted sum of chi squares. *National Aeronautics and Space Administration (NASA).* https://ntrs.nasa.gov/api/citations/19680015093/downloads/19680015093.pdf

Fel, T., Vigouroux, D., Cadène, R., & Serre, T. (2020). How good is your explanation? Algorithmic stability measures to assess the quality of explanations for deep neural networks. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1565–1575. Retrieved from https://api.semanticscholar.org/CorpusID:235727808

Fort, S., Hu, H., & Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *ArXiv*, *abs/1912.02757*. https://api.semanticscholar.org/CorpusID:208637294

Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1225–1234). https://proceedings.mlr.press/v48/hardt16.html

Higham, N. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and Its Applications, 103*(May 1988), 103-118. https://www.sciencedirect.com/science/article/pii/0024379588902236?via%3Dihub

Kearns, M., & Ron, D. (1997). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, *11*, 1427–1453. https://api.semanticscholar.org/CorpusID:6601319

Lee, J., & AlRegib, G. (2020). Gradients as a measure of uncertainty in neural networks. *2020 IEEE International Conference on Image Processing (ICIP)*, 2416–2420. doi:10.1109/ICIP40778.2020.9190679

Malmström, M., Skog, I., Axehill, D., & Gustafsson, F. (2020). Asymptotic prediction error variance for feedforward neural networks. *IFAC-PapersOnLine*, *53*(2), 1108–1113. doi:10.1016/j.ifacol.2020.12.1310

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 427–436. doi:10.1109/CVPR.2015.7298640

Russell, K., Green, C., Etheredge, C., Yohe, M., Marx, W., Hill, T., Odom, L., Drown, D. (2023). Using AI to increase trust in AI - yes, we're serious. *I/ITSEC 2023*. https://s3.amazonaws.com/amz.xcdsystem.com/44ECEE4F-033C-295C-BAE73278B7F9CA1D_abstract_File17275/FinalPaper_23316_0822051122.pdf

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. doi:10.1109/ICCV.2017.74

Uchida, S., & Sakoe, H. (1998). A monotonic and continuous two-dimensional warping based on dynamic programming. *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, *1*, 521–524 vol.1. https://api.semanticscholar.org/CorpusID:12238614