Using AI to Increase Trust in AI - Yes, We're Serious

Kyle Russell, Connor Green, Charles Etheredge, Michael Yohe,
William Marx, Ph.D., CAPT Timothy Hill, USN, (ret), Lt. Col (ret)
Lane Odom USAF, COL (ret) Daron Drown USAF
Intuitive Research and Technology Corporation
Internal Research and Development
Huntsville, AL
kyle.russell@irtc-hq.com; chad.etheredge@irtc-hq.com;
connor.green@irtc-hq.com; michael.yohe@irtc-hq.com;
william.marx@irtc-hq.com; timothy.hill@irtc-hq.com,
lane.odom@irtc-hq.com, daron.drown@irtc-hq.com

ABSTRACT

The use of Artificial Intelligence (AI) in high-consequence decision-making tasks presents legal, moral, and ethical challenges. Complicating the issue, AI systems have become more complex and are less explainable than deterministic or rules-based systems. AI agent behaviors are viewed as "black boxes," as their decisions can seem arbitrary or opaque to users, and in some cases, even for the developers of the system. Additionally, algorithmic performance is not guaranteed as complexity and number of inputs grows so does the difficulty in covering all corner cases. This lack of trust has significant negative consequences for the adoption of AI in decision-making.
To remedy this, the trust gap must be addressed by adapting the systems of verification and validation used for military technologies, particularly the Test and Evaluation capabilities. This includes developing individual testable metrics from contract requirements. This requires breaking apart functionality into small testable elements, with objective figures of merit, to make incremental improvements.

To advance the state-of-the art in establishing trust in AI for complex aerospace systems, an approach is proposed that is called "Digital Neurology." This approach can provide a real-time in-situ, trained, observer "agent" that can monitor and evaluate AI-based Neural Networks (NN). This paper will describe a concept for experimentation with autonomous aircraft. The observation agent would observe and infer behaviors that are indicative of:

- Normal and abnormal computational processing
- Anticipated and unexpected situations
- Regions of high and low computational use
- Regions of importance and insignificance

Similar to a human instructor following a human pilot through the training experience, an AI observer agent would be deployed with the autonomous aircraft systems, following and learning behavior to build trust. This paper presents an approach to build Trust in AI using this method.

ABOUT THE AUTHORS

Mr. Kyle Russell is a Senior Digital Engineer with Intuitive Research and Technology Corporation (*INTUITIVE*®). As a member of the research and development team he is responsible for exploring new applications of cutting-edge technologies to customer problems. He received a BS in Electrical Engineering from the University of Alabama and is currently pursuing a MS in Computer Science with focuses on Artificial Intelligence/Machine Learning and Data Analytics. He has experience developing advanced real-time interactive visualization solutions, and Big Data Analytics pipelines, as well as experience with modern web technologies and database systems in general.

Mr. Connor Green is a Senior Digital Engineer at *INTUITIVE*. As a member of the research and development team, he is responsible for staying current on the latest breakthroughs and techniques in deep learning to find interesting opportunities for customer solutions. He received a BS in Electrical Engineering from Mississippi State University and is currently pursuing an MS in Computer Science with a specialization in Artificial Intelligence and Machine

Learning. He has experience in RF seeker technology, Hardware-in-the-loop simulations, RF test chambers and direct signal injection labs, 3D printing, Long-Range Precision Fires, and deep learning neural networks.

Mr. Charles Etheredge is a Software Engineer with *INTUITIVE*. As a member of the research and development team, he is tasked with performing and demonstrating research for the company. He has received BS degrees in Computer Science and Business Administration from the University of North Alabama. He has expertise with video and metadata encoding solutions, as well as web, cloud, and AI applications. Additionally, he has participated in a wide variety of research, including data visualization, image analysis, audio analysis, cryptography, and AI weather forecasting.

Mr. Michael Yohe is the Simulation and Visualization Solutions Program Manager at *INTUITIVE* within the Software, Cyber, and Intel Division. He leads a team of engineers, developers, and artists in the development of data analytical tools for big data projects for business operations, cybersecurity, healthcare, and supply chain as a part of *INTUITIVE*'s independent research and development (IR&D) program. His background is software development and cybersecurity in software systems within private industry and defense for nearly twenty-five years and attained multiple patents for work in extended reality, big data analytics, and artificial intelligence. He received his BS in Information Technology from Oakwood University. He holds the CISSP certification.

Dr. William Marx is the Senior Vice President and Chief Technology Officer of *INTUITIVE*. He is responsible for planning, managing, and executing research and development programs aligned with the technology priorities of the US 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 with *INTUITIVE*. He is responsible for efficient operation of the Central Florida office and 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.

Lt. Col (ret) Lane Odom USAF has over 23 years of experience in aerospace technology development and is currently the Deputy Division Manager for Florida Operations at *INTUITIVE*'s Niceville, FL Office. Lane is a USAF Test Pilot School graduate with over 2500 flight hours in 31 unique aircraft types and served as Commander of the 586th Flight Test Squadron at Holloman AFB, NM. During his flight test career, he evaluated new technologies used for fighter aircraft, advanced weapons, sensors, and C2 Datalinks. At *INTUITIVE*, Lane led the team responsible for standing up initial operations of an autonomous aircraft testbed at Eglin AFB, the XQ-58 Valkyrie. Lane received his BS in Aerospace Engineering from Auburn University, and his MS in Aerospace Engineering from North Carolina State University.

COL (ret) Daron Drown USAF is a technical program manager and test and evaluation and autonomy SME at *INTUITIVE*. He is a retired Air Force Test Pilot with specific expertise in fifth generation combat aircraft, radar and optical sensors, guided munitions, AI-driven autonomous battle management, mission planning, and missile defense. He also has significant experience operating advanced military simulators and training systems, both as student and instructor. He earned a BS in Mechanical Engineering from the U.S. Air Force Academy, an MS in Engineering Mechanics from Michigan State University, and MS in Systems Engineering from the Air Force Inst. of Technology, and an MS in National Resource Strategy from National Defense University.

Using AI to Increase Trust in AI - Yes, We're Serious

Kyle Russell, Connor Green, Charles Etheredge, Michael Yohe,
William Marx, Ph.D., CAPT Timothy Hill, USN, (ret), Lt. Col (ret)
Lane Odom USAF, COL (ret) Daron Drown USAF
Intuitive Research and Technology Corporation
Internal Research and Development
Huntsville, AL
kyle.russell@irtc-hq.com; chad.etheredge@irtc-hq.com;
connor.green@irtc-hq.com; michael.yohe@irtc-hq.com;
william.marx@irtc-hq.com; timothy.hill@irtc-hq.com,
lane.odom@irtc-hq.com, daron.drown@irtc-hq.com

INTRODUCTION

The use of Artificial Intelligence (AI) in high-consequence decision-making tasks presents legal, moral, and ethical challenges. Complicating the issue, AI systems have become more complex and are less explainable than deterministic or rules-based systems. AI agent behaviors are viewed as a black box, as their decisions can seem arbitrary or opaque to users, and in some cases, even for the developers of the system. As algorithmic complexity and number of inputs grows, so does the difficulty in covering all corner cases. The increase of complexity affects the number of potential outcomes or decisions of the AI agent. Therefore, it can be difficult to fully test an AI agent, potentially leading to a lack of trust and a resulting decrease in the adoption rate of AI in decision-making.
Before widespread adoption of AI technology can occur, the trust gap must be addressed. To accomplish this, the verification and validation methods already used for military systems and technologies can be adapted to the AI domain. As part of this process, individual testable metrics will need to be derived from the contract requirements. This derivation requires breaking apart functionality into small testable elements, with objective figures of merit, to make incremental improvements.

To advance the state-of-the art in establishing trust in AI for complex aerospace systems, “Digital Neurology” is an approach that we propose that can provide a real-time in-situ, trained, observer “agent” that can monitor and evaluate AI-based Neural Networks (NN). The observation agent would observe and infer behaviors that are indicative of:

- Normal and abnormal computational processing
- Anticipated and unexpected situations
- Regions of high and low computational use
- Regions of importance and insignificance

Of particular significance for the Defense Training and Simulation Community is the ability to train alongside these AI systems and trust that the decision-making behavior developed and exhibited during training will continue during deployment, much the same as a well-trained person. With this concept, it is postulated that the observer agent will engender more trust in the original AI agent by identifying behaviors observed in the AI agent during specific conditions and scenarios. It can also potentially shed light on why the AI agent exhibited those behaviors in response to environments and stimuli. The insights provided by the observer agent will allow humans to have greater understanding into the innerworkings of the AI agent by contributing data, context, and information to support increasing human trust of the AI systems by moving the AI agent further away from the black box category.

PRIOR RESEARCH

Trust in AI is a crucial factor for the adoption and acceptance of AI technologies in various domains. However, trust in AI is not a well-defined concept and lacks a common theoretical and methodological foundation. In this section, recent literature will be briefly reviewed on trust in AI, highlighting some of the key challenges and opportunities for research.

Fundamental Trust In AI

One of the challenges for trust research in AI is establishing a clear definition and measurement of trust in AI. According to Lukyanenko et al. (2022), trust in AI is a problem of interaction among systems and can be understood using systems thinking and general systems theory. They propose a framework that provides a basis for trust research in general and trust in AI specifically. The framework distinguishes between different types of systems (e.g., natural, artificial, social) and different levels of analysis (e.g., individual, group, organizational), and suggests trust can be examined from multiple perspectives (e.g., cognitive, affective, behavioral).

Another challenge is understanding how aspects of AI, such as transparency and interpretability, affect trust formation and maintenance. Transparency refers to the degree to which an AI system reveals its inner workings, logic, and rationale to its users or stakeholders. Interpretability is the ability of an AI system to provide understandable and meaningful explanations for its actions or decisions. According to Kizilcec et al. (2021), transparency and interpretability are not synonymous and have different effects on trust depending on the context and the user's goals. They suggest that transparency can increase trust by reducing uncertainty, enhancing accountability, and fostering learning, while interpretability can increase trust by improving understanding, enabling feedback, and supporting collaboration.

Once established, maintaining trust in AI must be explored to understand the factors contributing to aversion and distrust of AI, negative attitudes or behaviors toward AI systems that hinder their adoption. According to Baker and Wurgler (2007), people tend to show a bias for human judgment and build an aversion to algorithmic judgment when observing mistakes. In their study, they found that participants seeing a model make relatively small mistakes consistently decreased confidence in the AI model, whereas seeing a human make relatively large mistakes did not consistently decrease confidence in the human. They argue this is due to the difficulty humans have in evaluating the quality of algorithmic output compared to human output and the lack of ground truth on which to measure or evaluate the results of learned algorithms. They also suggest that people are more forgiving of human errors because they attribute human errors to situational factors, while attributing algorithmic errors to dispositional factors. A later study by Glickson (2021) exploring trust in human-human vs human-AI interaction appears to strengthen the findings of the Baker and Wurgler study. Glickson found that for human-human interaction, human trust generally starts low, but tends to build over time and with experience; for human-AI interactions, high initially trust occurs, but this trust typically trends downward with experience.

Human Decision-Making Based on AI

A human's confidence in decision-making relies on their perceived understanding and risk-management capabilities. An ideal decision-maker would have substantial domain-specific education and experience, a history of success using the current decision-making pipeline, a suite of useful metrics to observe, and a counsel of peers. A lack in any of these areas pushes the decision-maker to rely more heavily on the others; thus, increasing risk as the final decision strays closer to a single point of failure, and the decision-maker's trust in their own decision decreases. In a human and AI team, the human does not need to trust the AI's output completely if the human is confident that any mistakes can be caught by the rest of the decision-making pipeline. This is similar to a decision-maker knowing when to trust or disregard the opinion of a peer without being able to see into the peer’s brain (Esposito, 2021). Understanding one’s peers, in the context of domain specific decision-making, relies on the shared experience of being a human in that industry; more specifically, understanding the reasoning that stems from various factors such as common background for education, possible bias, and priorities.

In a medical publication on clinical decision-making, Giordano (2021) examined the requisites and limitations faced by physicians when integrating AI applications. Their results indicate trust can be built around the black box nature of AI models with proper physician training and communication from the model. In particular, application bias, where a model exhibits bias in a subset of predictions due to bias in its training data, requires physicians to be vigilant and educated on the model when applying its results to patients. Physicians need to be aware that some AI results may not align with the physician’s priorities for a patient; an example being that an AI utilizing a heuristic reward function might incorrectly favor short term effects, when the physician is trying to treat a chronic affliction for an atypical patient. This concept can be applied to AI models in other domains as well.

Observation and Transparency

An AI's role in a team can be that of a peer or as a source of dynamically curated metrics for the decision-maker; in either role, the AI team member's decisions need some level of transparency before the human decision-maker can feel comfortable incorporating them into the decision-making process (Yu & Li, 2021). Transparency is not a replacement for thorough model explanation, but it can act as a substitute; this is often achieved through education. By providing users with information explaining a model's training process and performance history across a wide verity of use cases, users will build a higher level of trust in the model (Esposito, 2021). In addition, humans focus more on critical feedback, thus trust in AI can be improved if a human believes they have seen most of the ways an AI can make a mistake, and which indicators to look out for that signal a low trust decision. In a study performed on AI moderating online content, it was found that trust was improved when humans were able to provide feedback to the AI and have a sense of agency in the decisions it made (Molina, 2022).

OUR RESEARCH

Though the human brain and a neural network are different; they can still be thought of as similar in many aspects. Both the human brain and a neural network make decisions based on a series of inputs or stimuli. Because of this, it is hypothesized that some traditional medical neurology approaches may be translatable to understanding how AI brains perform and what makes them behave in certain ways. This is the fundamental concept for the "Digital Neurology" approach for creating trust in AI.

In Figure 1, a high-level concept for Digital Neurology is illustrated. There are three levels to the process. The first is a behavioral observation approach; this is important in human pathology, as many diagnoses can only be made via analysis of patient behaviors and symptoms. This parallels AI development as behavior can drastically change during training, even among cohorts that share the same base model.

The second level is a more direct, metric driven approach. In humans it consists of measuring electrical signals to detect brainwaves and abnormal brain activity. Parallels can be drawn with AI by stimulating the network with an input and recording the output, building up a complete map of its decision space.

The third level of Digital Neurology draws on the modern techniques of functional-MRI. For human patients this not only allows us to see the signals of the brain, but specifically where these signals are occurring in the 3D brain topology, thereby providing insight into the structure and function of the brain. For AI, it is postulated that this could also be the case with certain portions of the network being responsible for certain functions, and this information would ultimately lead to higher trust AI systems.

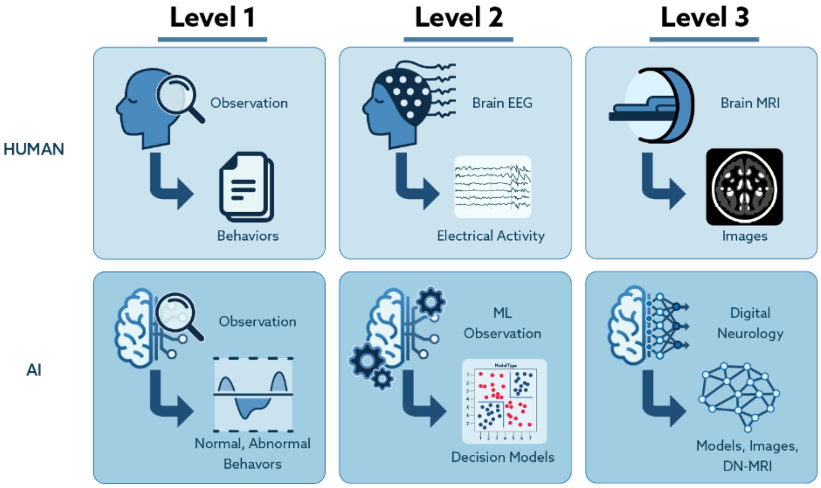

Figure 1. Visual Depiction of How Digital Neurology Relates to Human Neurology

This paper covers the first two levels of Digital Neurology and some elements of the third as shown in Figure 1 above. The first level consists of behavioral observation and related metrics. Typical metrics one might use to evaluate performance would consist of accuracy, precision, and recall; or if looking for a single metric the F1 score which relies upon precision and recall. Metrics, such as these, can give an overall impression of the performance of a model; however, they require interpretation by analysts and really do nothing to explain the behaviors of the model in question. The second level is a slightly deeper dive into analyzing the observed behavior of the model. Analysis is performed on outputs from the model to glean explanations for behavior in relation to the input. The third level is invasive analysis of a model's inner workings to attempt to attach meaning to specific areas in the network.

THE SYSTEM

The proposed system is based on peer research, internal research, and development activities. For the system, two models will be used: a black box model, and an observer clone model. Behavioral cloning will be used to duplicate the input-output mapping of the black box model into the more interpretable observer model. The authors have previously shown the effectiveness of a behavioral cloning approach using the eXtreme Gradient Boosting technique (XGBoost). In the previous work, 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). Creating a faithful clone to the original model allows the usage for additional interpretability techniques by exposing internal back-propagation data. Figure 2 is visualization for the planned pipeline. Unlike the black box model, the input to the cloned observer model will be supplemented with the analysis results and will be more transparent due to its exposed layers. For this approach, a black box model should be selected that provides access to its layer metrics to help verify that the cloned model is faithful enough to the original model.

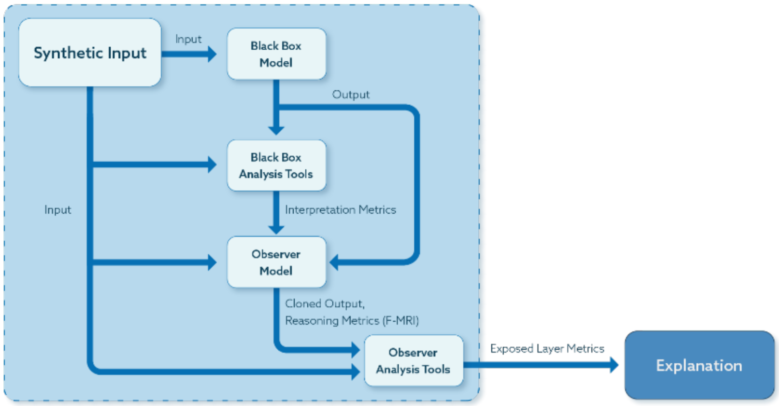

Figure 2. A Graphical Overview of the Approach

This approach recreates the common scenario in which a black box AI model is delivered without its original training data. In place of the original training data, a synthetic dataset can be generated for collecting the black box model's output. The new input and output will undergo analysis using techniques and libraries to create a more extensive dataset to feed the observer model as it tries to mimic the black box model with faithful reasoning; to achieve this, feature-to-decision causality needs to be identified. This will necessitate an environment that can not only generate test data, but also manipulate a copy of the test data to create counterfactual examples; in the case of driving data, this could be demonstrated by photoshopping pedestrians, lighting changes, or a noisy sensor to find the bare minimum needed to induce avoidant steering. Useful input perturbations need to be coherent and sparse while still managing to drastically affect the output. This can be automated via a generative adversarial network. However, for this initial approach, it is more likely to be done by hand at first. In addition to adversarial testing, the transparency of the observer model and training environment flexibility will allow for validating results with various libraries and thus drive improvements.

The Black Box Model

The black box model will be a representative analogue for a generic aviation control system. In the case of this approach, NVIDIA's PilotNet self-driving model can be used as the black box model (Huang, 2017). As seen in Figure 3, PilotNet takes vehicle camera data and predicts the steering wheel angle. This model was chosen due to its widely applicable control objectives, built-in visualization of layer reasoning, and ease of data generation.

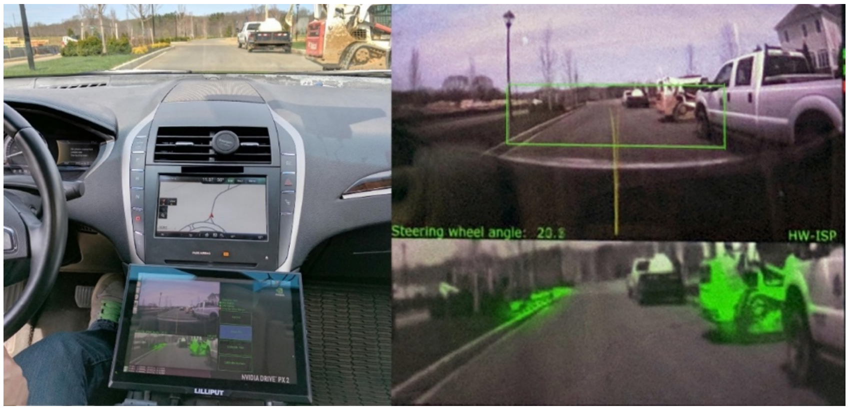

Figure 3: Example of PilotNet in operation. Using an internal feature of PilotNet, the pixels in green are salient features that has the model's focus and will be used as a faithfulness check against the cloned observer model's attention (NVIDIA, 2016).

Generating Synthetic Input Data

When developing a dataset, special considerations are needed for using behavioral cloning to improve trust in a black box model. High volumes of training samples are needed to train the cloned observer model; in addition, attempts at automating the sample generation will be complicated by the additional need for sample variety. However, the virtual nature of a simulator environment allows for the utilization of reinforcement learning and other automation techniques; thus, it's possible to generate datasets with more problem space coverage than real world data collection. This is largely due to the relative infrequency of real-word examples that are crucial to learned behavior. In the case of autonomous driving, important edge cases could include traffic accidents, faulty signaling, blinding sun glare, or other environmental parameters that affect behavior while being underrepresented in the bulk of traffic data. These edge cases are relatively easy to include in synthetic data. By using testing effectiveness as a policy, reinforcement learning can be used to generate scenarios until the experiment's volume and variety needs are met.

This approach to synthetic data generation is appealing because it is comparable to a scenario where a contractor delivers a flight model without training data to a customer. If the customer or verification 3rd party lacked their own test data, then a flight sim could act as a source of rapid synthetic data generation. Even if the visual fidelity is not on par with the real-world data, the simpler graphics can be a boon by identifying proper causal relationships while minimizing spurious correlations when evaluating a model's decision-making.

The built-in visualization of the PilotNet model will help verify that the observer model is faithful to the original AI's reasoning. Also, since the input is video based, new input data can be generated manually through dash cam footage paired with steering angle, or more interestingly, through a modern video game engine. A readily available solution is an open-source simulator for autonomous driving research known as CARLA (Car Learning to Act). CARLA was used to great effect in the original paper discussing the technique (Ruiz et al., 2019). The features CARLA already contains and its support from NVIDIA would significantly reduce environment development time and provide a more robust set of test scenarios. A sample of CARLA's visuals can be seen in Figure 4.

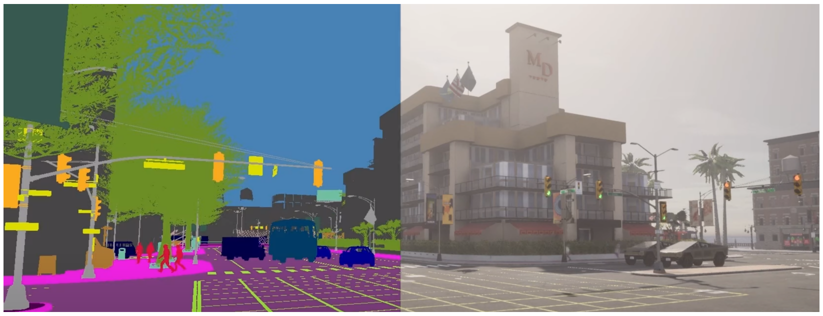

Figure 4: Example of environment in CARLA simulator. The left side demonstrates the semantic visual filter feature, and the right side demonstrates its typical appearance for sunny weather (CARLA, n.d.).

The Observer Model

After a training dataset has been created, the observer model can be created and trained. A Transformer model has been chosen due to the continuous nature of a control system, its attention mechanism, and the enhanced explainability inherent to the architecture. Attention mechanisms improve model interpretability because they allow the observing model to learn the observations that are critical to the outcome (Vaswani et al., 2017). These mechanisms imitate human intuition by seeking out the areas of a given input that impact a prediction the most, similar to how a driver may focus more on the adjacent lane when merging lanes, while lowering the priority to the distance of the car ahead.

After the observer model's architecture has been established, the dataset created using the black box model can be used to train the observer model. The training process is a standard supervised learning procedure, which means the black box dataset is split between input and label, the input is then fed to the model, and the output of the model is recorded and compared to the label (ground-truth).

During training, it is imperative that the model loss be evaluated; in general, the goal is to reduce the loss of the observer model predictions to match those of the black box model. Just like any other model, care is needed to balance performance with overfitting. Additionally, the reasoning behind the predictions needs to match the black box, but this is not a metric that would be available for most black box models. Using the post-hoc interpretability methods described below, a level of reasoning can be gleamed when monitoring the output of a black box model while the input is varied. It is then hypothesized that supplementing the input with the analysis results could aid in shaping the reasoning of the cloned observer model to better fit the black box model. As shown in Figure 5, the hypothesis can then be evaluated by comparing the observer's transparent layers to the black box model's internal visualization; normally a black box model may not have this feature, but the selected model for this experiment does. If the experiment is successful, then this approach could still be applied to black box models that lack a partial glimpse into its inner workings.

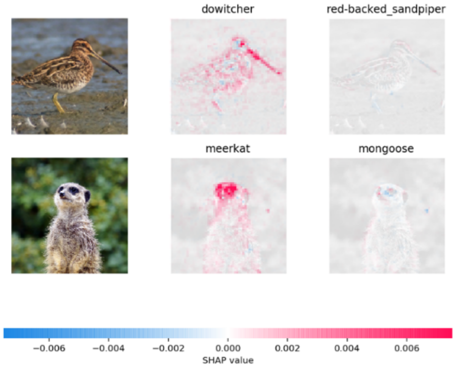

Figure 5. Example of SHAP library output of a classification explainer where features are colored based on its impact towards the decision (Lundberg, 2021).

Interpretation Analysis Tools

Comparison between both models will require an evaluation method that is not specific to any one model type. Shapley values are a model-agnostic technique that facilitates direct comparison between different model types. Shapley values require only black box access to the model as a predictor and can be computed without any knowledge of the model's internal mechanics. This enables comparison between a trained observer and subordinate model. As shown in Figure 5, these values measure the relative impact of input features on output by comparing an average marginal contribution of each input feature to the overall model score.

The Shapley value for a given feature is calculated using the formula shown in Figure 6. In this equation, |z`| is the number of non-zero entries in z`, and z` ⊆ x` represents all z` vectors where the non-zero entries are a subset of the non-zero entries in x`.

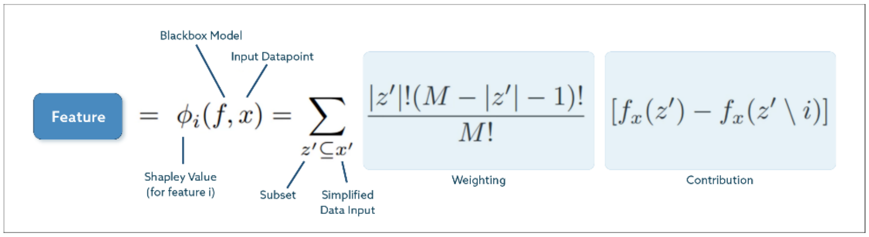
Blackbox Model
Input Datapoint
Feature
$= \phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$
Shapley Value (for feature i)
Subset
Simplified Data Input
Weighting
Contribution

Figure 6. The formula for calculating the Shapley value for a given feature.

The average model score used when computing Shapley values represents the average output of a model. For example, if a model predicts the remaining lifespan of a component, its average score would be the average length of time predicted across all datapoints in a dataset. This score is typically calculated using the dataset that was used to train the model. This calculation is computationally expensive, but a close approximation can be computed using the Shapley Additive Explanation (SHAP) approach and a randomly selected subset of datapoints (Molnar & Kursawe, 2017). The proposed approach goes a step beyond this, using synthetic data that has been generated by other means, such as the simulation method described above. After an average model score is determined, the impact of a given feature can be calculated in three main steps. The first is to select a random input datapoint from the dataset and record the output for this given input. Next, the input parameter of only the feature in question should be modified to its corresponding average for the model score and the result of this input should be recorded. Finally, by comparing the results of the sample input and the modified input, a Shapley value representing the impact this feature contributes to the average score for this sample is created. This process can then be repeated for many samples across all features to determine the average marginal contribution of each feature in the model.

Reasoning Tools

During the back-propagation step of an AI, it is possible to capture details at the layer level of a model; using these details, there are several tools that can visualize a model's reasoning (Zablocki, 2021). Unless the black box is designed with internal hooks, this type of method would be impossible. However, behavior cloning into a more transparent model would allow the use of propagation-based tools. As long as the behavior cloned model follows the same prediction and reasoning patterns as the original, it can be used in place of the black box to help an analyst provide an explanation for a prediction. For convolutional networks, like PilotNet, a technique known as Grad-CAM can be used to visualize an internal pattern for a given input-output pair. For transformers like the cloned observer model, an example of a similar technique would be the Transformer-Explainability library by Chefer et al. (2020). Figure 7 demonstrates a typical output from these types of tools.

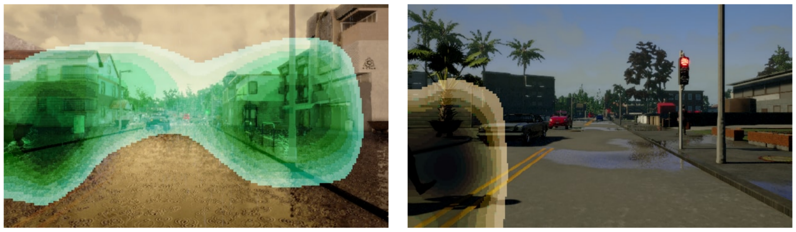

Figure 7: Representative example of visualizing a model's reasoning (Gupta et al., 2018).

The Result

One facet of improving trust in an AI black box model is making its decision-making more explainable. Despite the lack of transparency, there are methods to interpret a model's reasoning with only input perturbations and adversarial testing. Propagation-based methods become available when another, more transparent, AI is introduced through behavior cloning. However, a behavior cloned observer model may achieve the same decisions, but it does not guarantee the same reasoning as the original model. This paper proposes that supplementing the input to the clone with interpretability metrics will help it to better mimic the original's reasoning; thus, allowing the model to better act as dissectible surrogate to the black box model. Confirming this hypothesis will be accomplished by demonstrating that both the predictions and reasoning patterns accurately match without overfitting for a comprehensive set of test cases. For example, if both models predict a steering angle of 10 degrees, then it would have its reasoning justified if both models had similar focus patterns across the camera video.

In summary, our proposed approach for Digital Neurology to increase trust in AI can be described by the following sequence of steps:

1. Employ the CARLA simulator to automatically create enough synthetic camera footage to meet volume and variety requirements.
2. Record outputs from the PilotNet which is acting as the black box analogue.
3. Generate interpretation metrics from PilotNet outputs using various tools such as the SHAP library.
4. Train the observer model to match the black box mode's output using the footage input and interpretation metrics generated previously.
5. Compare the reasoning patterns between the observer and black box once training has converged.
6. Validate hypothesis by comparing the observer reasoning with the black box reasoning using PilotNet's internal visualization capability (only for this proof).
7. Generate an explanation report (manually) from the results of the interpretability tools on the black box model and additional reasoning metrics from the observer model.

CONCLUSIONS AND FUTURE WORK

Complex, decision-making, critical AI agents are often perceived with skepticism as these applications are typically harder to test and evaluate due to larger problem spaces. These AI agents are considered black boxes because the rationale for why they make certain decisions is not readily apparent to the user and, in some cases, the developers. Many of these AI agents present legal, moral, and ethical challenges, making the need for comprehensive training and evaluation of these agents critical.

In order to remedy the lack of trust in a black box AI, traditional verification, validation, and test capabilities employed for testing military systems can be adapted to test the AI. The system postulated in this paper will begin to address this trust gap by allowing testers to behaviorally clone a black box AI agent; thereby, improving the evaluation of the innerworkings and rationale behind an AI agent's decisions. In addition, if this level of insight is provided over multiple projects, analysts will begin to develop an intuition for a trustworthy baseline for managing future AI-based projects. With success in the presented concept, an observer agent would establish more trust in the observed agent by highlighting the reasoning behind behaviors during specific conditions and scenarios.

In this paper, the proposed system focuses on the evaluation of autonomous vehicles. However, this system and its testing can be applied to a multitude of different AI agents in the military domain including military control systems and autonomous aircraft. Additionally, further development of evaluation agents could result in automation of the review process with minimal, but necessary, oversight from human analysts. Using a dataset built from the analysis results of subject matter experts, the metrics that correspond most strongly with trust could be trained into an AI agent for it to interpret trustworthiness in a reliable and consistent way; thus, preventing knowledge and skill erosion as it becomes a maintained system that inexperienced analysts could learn from. Further work might also entail advanced adversarial testing for detection of subtle discrepancies and resultant behaviors. Breakthroughs in AI technology are happening on an almost yearly basis, and performance increases follow Moore's law due to hardware progress; thus, it is imperative to maintain awareness on peer research and cutting-edge solutions in order to remain competitive on global scale as new possible applications become available.

REFERENCES

Baker, M. & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129-151. **https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce_papers**

CARLA. (n.d.). CARLA - An open urban driving simulator. **http://carla.org/**

Chefer, H., Gur, S., & Wolf, L. (2020). Transformer interpretability beyond attention visualization. **https://arxiv.org/pdf/2012.09838v1.pdf**

Esposito, E. (2021). Transparency versus explanation: the role of ambiguity in legal AI. *Journal of Cross-Disciplinary Research in Computational Law 1* (2). **https://journalcrcl.org/crcl/article/view/10**

Etheredge, C., Russell, K., Marx, W., Hill, T. & Drown, D. (2022). The use of AI/ML to replicate threat behavior for nonlinear simulation. In Proceedings of I/ITSEC (pp. 1-10).

Giordano C., Brennan M., Mohamed B., Rashidi P., Modave F., & Tighe P. Accessing artificial intelligence for clinical decision-making. *Front Digit Health.* 2021;3:645232. doi:10.3389/fdgth.2021.645232

Glickson, J. & Keil, F. (2021). Human Trust in Artificial Intelligence: Review of Empirical Research. *Annual Review of Organizational Psychology and Organizational Behavior, 8*(1), 1-25. doi:10.5465/annals.2018.0057

Gupta, S., Davidson, J., Levine, S., Sukthankar, R., & Malik, J. (2018). Conditional Affordance Learning for Driving in Urban Environments. **https://arxiv.org/abs/1806.06498**

Huang, L. (2017, May 10). *PilotNet*. GitHub. **https://github.com/lhzlhz/PilotNet**

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., & Schneider, E. (2021). Explaining why the computer says no: algorithmic transparency affects perceived justice and willingness to use decision support systems. *Public Administration Review, 81*(2), 299-309. **https://onlinelibrary.wiley.com/doi/10.1111/puar.13483**

Lukyanenko, R., Maass, W. & Storey, V.C. (2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electron Markets 32*, 1993–2020. **https://doi.org/10.1007/s12525-022-00605-4.**

Lundberg, S. (2021, May 14). *SHAP (SHapley Additive exPlanations)*. GitHub. **https://github.com/slundberg/shap**

Molina, M. & Sundar, S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication, Volume 27* (4). **https://doi.org/10.1093/jcmc/zmac010**

Molnar, C., & Kursawe, K. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30* (NIPS 2017), 4765-4774.

NVIDIA. (2016, June 8). *Explaining How End-to-End Deep Learning Steers a Self-Driving Car*. NVIDIA Developer. **https://developer.nvidia.com/blog/explaining-deep-learning-self-driving-car/**

Ruiz, N., Schulter, S., & Chandraker, M. (2019). Learning to simulate. arXiv arXiv:1810.02513v2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Yu, L. & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: the parallel multiple mediating effect of effectiveness and discomfort. *Behav Sci (Basel), 12*(5):127. doi:10.3390/bs12050127

Zablocki, E., Ben-Younes, H., Perez, P., & Cord, M. (2021). Explainability of deep vision-based autonomous driving systems trained with behavior cloning. arXiv preprint arXiv:2101.05307.