Fostering Trust and Quantifying Value of AI and ML

Dalmo Cirne
Machine Learning for Financials
Workday
Boulder, Colorado, USA
email: dalmo.cirne@workday.com

Veena Calambur
Responsible AI
Workday
Princeton, New Jersey, USA
veena.calambur@workday.com

***Abstract*—Artificial Intelligence (AI) and Machine Learning (ML) providers have a responsibility to develop valid and reliable systems. Much has been discussed about trusting AI and ML inferences (the process of running live data through a trained AI model to make a prediction or solve a task), but little has been done to define what that means. Those in the space of ML-based products are familiar with topics such as transparency, explainability, safety, bias, and so forth. Yet, there are no frameworks to quantify and measure those. Producing ever more trustworthy machine learning inferences is a path to increase the value of products (i.e., increased trust in the results) and to engage in conversations with users to gather feedback to improve products. In this paper, we begin by examining the dynamic of trust between a provider (Trustor) and users (Trustees). Trustors are required to be trusting and trustworthy, whereas trustees need not be trusting nor trustworthy. The challenge for trustors is to provide results that are good enough to make a trustee increase their level of trust above a minimum threshold for: 1- doing business together; 2- continuation of service. We conclude by defining and proposing a framework, and a set of viable metrics, to be used for computing a *trust score* and objectively understand how trustworthy a machine learning system can claim to be, plus their behavior over time.**

Keywords-artificial intelligence, machine learning, trust, game theory.

I. Introduction

Much has been said about responsible Artificial Intelligence (AI), but the majority of those conversations are high-level and focused on defining principles—which are important for defining direction—but are rarely coupled with the actual operation of ML-based systems.

Measuring the increase or decrease of trust in this technology is a gap that needs to be addressed, and that is the main proposal of this paper: a quantitative framework to be used in computing the trustworthiness of AI and ML systems. Here, trust is defined as the willingness to interact with an AI/ML system while being aware that a model inference [1] is fallible.

The framework, however, is not without its challenges. There are several other elements to be considered in an AI/ML-powered system in order for it to gain the trust of its users. Good inferences are one of them, but so is data privacy, mitigating bias, measuring qualitative aspects, tracking the trust level over time, model training automation, and so on.

The paradigm explored in this paper assumes that trust is built by the trustor's initial act, signaling that the actor is trustworthy. More specifically, the trustor's act would be to invest in building a product and offer it to customers with the promise that it will generate value to them; more value than what is paid in return for the service. The trustor decides how much to invest, and the trustee decides whether to reciprocate and give continuity to the business relationship.

Note that the trustee does not have to be held to similar standards for trustworthiness as the trustor. The objective is to make the customers trusting—above a minimum threshold T—as to engage in the *Trust Games* [2]. These games are extensions built on top of the *Game Theory* [3]. Furthermore, trust has a temporal element to it. Once established, there are no guarantees that there will be a continuation. Therefore, this is an extensive form of interaction where both actors collaborate and observe each other, reacting to historical actions from one another.

A global study, conducted by the services and consulting firm KPMG, and named "Trust in Artificial Intelligence [4]," has found that there is a wariness sentiment in large sections of the workforce in general. The people surveyed in the study expressed concern about trusting those systems, from financials to human capital management products. The framework proposed in this paper will help address such sentiment by quantifying and measuring trust in AI and ML. The results can then be shared with the workforce or the population as a whole to help them better understand how ML-based solutions function and in turn, develop a positive sentiment towards adopting such products.

The rest of the paper is structured as follows. In Section II, we examine the dynamic of trust between a provider (Trustor) and users (Trustees). In Section III, we propose a quantification of trust over many iterations between trustor and trustee. In Section IV, we define a minimum trust threshold. In Section V, we present simulations of the quantification of trust. In Section VI, we present the categories for measuring trust. In Section VII, we demonstrate how the trust score can be practically implemented. In Section VIII, we define a region of fair trading between trustor and trustee. Section IX concludes our work.

II. Trust Games

The motion of a *trust game* is developed around two actors: a trustor and a trustee. The trustor has a service of value V to offer to a trustee. The value in question is *quality machine learning inferences*. ML is implemented as a software service, and by its nature, software can be replicated to any number n of customers without physical constraints. Thus, V can be offered independently and concurrently to all customers.

It could be the case that the value V of inferences may be only partially absorbed by a trustee. The limited, portioned consumption could be due to a variety of reasons, including, but not limited to: eligibility or capacity to use all the features (i.e., satisfies all requirements), service subscription tiers, users have yet to be trained.

In order to represent the range of scenarios where the trustor may transfer the entirety of value V or a smaller portion of it, we introduce a multiplier p, where $\{p \in \mathbb{R} \mid 0 \leq p \leq 1\}$. Therefore, the initial remittance sent by trustor u is:

$$R_u = pV \tag{1}$$

Depending on the quality of the trustor's results, trustees' perception of value may be magnified or reduced by a factor K, where $\{K \in \mathbb{R}\}$. For $K > 1$, it means that the trustor improved the efficiency of operations for the trustee (they do better than operating on their own). For $K = 1$, the trustee is operating at the same efficiency, and for $K < 1$ (negative values are also possible) the trustee is less efficient than before they started using the service. The initial perceived gain received by trustee v is:

$$\begin{aligned} G_v &= KR_u \\ &= KpV \end{aligned} \tag{2}$$

A trustee is free to reciprocate or not. During a trial period, they may choose to decline further service. Even if under contract, they may choose to skip renewal. On the other hand, assuming that the value received from ML inferences improved their efficiency, the incentive is to continue to engage. In either case, a trustee will give back a portion q of the gain received, where $\{q \in \mathbb{R} \mid 0 \leq q < 1\}$. The value sent back may take the form of monetary payment for the service, interviews, usability feedback, labeling of transactions, or a combination of those. The repayment B expected by trustor u is therefore:

$$\begin{aligned} B_u &= qG_v \\ &= qKpV \end{aligned} \tag{3}$$

There could be a consideration to introduce a magnification factor on the repayment from trustee v. That, however, is not necessary in the scope of this paper since trustees do not need to be trustworthy; the trustor u is not evaluating whether to trust them or not.

Fig. 1 represents the flow of the initial step in this trust game. The blue line segment represents the range of possible values delivered to trustees by the trustor, the large blue circle is the magnification factor applied to the value delivered, and the orange line segment represents the range of possible values reciprocated to the trustor by a trustee.

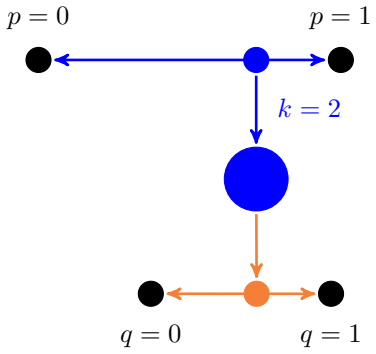


Figure 1. Trust Game payoffs.

Regarding the magnification factor, when $K > 1$, the value received back by trustor u is positive and enables the necessary conditions for an extensive form of the trust game (long-term engagement). It becomes a strong indicator that trustee v trustiness towards trustor u is equal or above the minimum threshold T, where $\{T \in \mathbb{R} \mid 0 \leq T \leq 1\}$.

When $0 \leq K < 1$, the service is causing the trustee some form of disruption (in the sense that efficiency has dropped below the level prior to using the service). This would be acceptable during the development phase of a product where the trustee takes part in a beta test program. In such a situation, the trustee sees a benefit in participating, assuming future value in adopting the service and the ability to harvest the benefits early on.

The worst-case scenario happens when $K < 0$. This could lead to rapid erosion of trustor u trustworthiness, customer churn, and other negative outcomes.

III. Quantifying Trust

The aim of this trust game is to create the circumstances necessary for repeated interactions between trustor and trustee.

After the initial remittance R_u, given by (1), there may be a residual value r on the trustor's side that a trustee did not take advantage of. For instance, maybe not all product features are being used, inference happens in batches and data is yet to be sent through the pipeline, or some other reason. That residual value is what is left from V:

$$\begin{aligned} r_u &= V - R_u \\ &= V - pV \\ &= (1-p)V \end{aligned} \tag{4}$$

The accumulated value A for trustor u upon completing the first cycle is the residual value r_u (4) plus the repayment B_u (3) received from the trustee:

$$\begin{aligned} A_u^{\text{1st cycle}} &= r_u^1 + B_u^1 \\ &= (1-p_1)V + q_1K_1p_1V \\ &= V(1 - p_1 + q_1K_1p_1) \end{aligned} \tag{5}$$

On the trustee's side, they will have received a value of G_v (2) and given back a portion q of it. The net gain N for trustee v at the end of the first cycle is:

$$N_v^{\text{1st cycle}} = G_v^1 - q_1 G_v^1 = (1 - q_1) K_1 p_1 V \quad (6)$$

Generalizing the gains for trustor and trustee for n cycles of the trust game, we have equations for trustor:

$$A_u = V \left(1 - \sum_{i=1}^{n} p_i + \sum_{i=1}^{n} (q_i) \sum_{i=1}^{n} (K_i) \sum_{i=1}^{n} (p_i) \right) \quad (7)$$

and trustee:

$$N_v = V \left(1 - \sum_{i=1}^{n} q_i \right) \sum_{i=1}^{n} (K_i) \sum_{i=1}^{n} (p_i) \quad (8)$$

The objective is to maximize the payoff to the trustee and trustor—possibly skewed towards the trustee. As such, trust has to be repaid [5] (i.e., $q > 0$). The trustor benefits from economies of scale by the aggregate of payoffs from all trustees.

IV. Threshold

For a trustor to increase its trustworthiness (W_u) in the eyes of a trustee, the gains delivered by the service must be higher than if the trustee was operating on their own. Such condition is satisfied by the following system of inequalities:

$$W_u \subseteq \begin{cases} pV \geq T \\ K \geq 1 \end{cases} \quad (9)$$

That happens when the value of the remittance R_u is equal or greater than the threshold T (the value sent is at a minimum equal to the perceived value received), and the magnification factor K is greater or equal to one.

Being a system of inequalities, it is also possible to have a lower remittance ($pV < T$) and increase trustworthiness, as long as the magnification factor is large enough ($K \gg 1$) to make up for the shortfall. Although plausible, this would be uncommon.

V. Simulations

The following is a set of four simulations testing scenarios from fostering to eroding trust as a result of the quality of machine learning inference.

All the simulations begin from the same exact starting point, where it is assumed that the potential value of a product being offered to customers is of one million points (1,000,000). The starting number is an arbitrary value and could have been any positive number. We want to observe the shape of the curve formed from plotting interaction cycle after interaction cycle.

The hypothesis is that a trustee would increase their trustiness level towards the trustor by providing good machine learning inferences. Conversely, less than good enough results would have the opposite effect (i.e., erode trust).

Notice that throughout all four simulations, all parameters are kept the same, varying only the magnification factor K.

A. Simulation 1: Machine Learning Inferences Add Value

For this simulation, we will go step-by-step in the first interaction. For subsequent simulations, only the final graph plots will be shown. Irrespective of the simulation, they all can be reproduced using the source code [6] that accompanies this paper.

Assume that in the first cycle iteration, the trustor begins with V = 1,000,000 points and is able to send a remittance of 65% ($R_u = 0.65 \times$1,000,000) of inference value to a trustee. The magnification factor perceived by the trustee is $K = 2$, thus, the gain becomes 1,300,000 ($G_v = 2 \times$650,000) points.

The trustee sends a portion ($q = 0.14$) of the value back by interacting with the user interface, providing a feedback label, and paying for the service. The rebate received by the trustor is 182,000 ($B_u = 0.14 \times$1,300,000) points.

Adding the rebate to the residual value ($r_u = 0.35 \times$1,000,000), the trustor's accumulated gain equals 532,000 (A_u = 350,000 + 182,000) points, and the trustee's gain would be 1,118,000 ($N_v = 0.86 \times$ 1,300,000) points.

First, the trustee's perception was that they received more value than what the trustor had to offer due to the magnification factor (win). Second, the trustor received a rebate in various formats—accruing value that was not there before (win). And third, after the aggregate across all trustees, the trustor will have accumulated more than the initial value offered (win).

In Fig. 2, we can see the shape of the curve showing the accumulated gains for both trustor and trustee for the four cycles of the simulation.

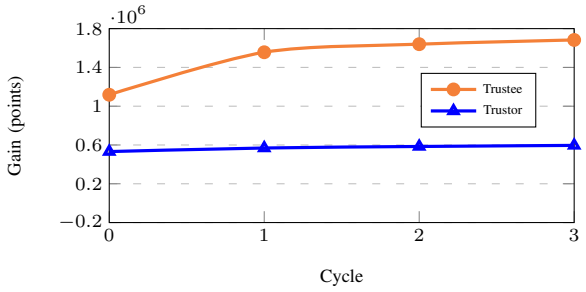


Figure 2. Accumulated gains ($K > 1$).

B. Simulation 2: Machine Learning Inferences Are Neutral

For the second simulation, a neutral magnification factor ($K = 1$) is being simulated. The value sent by the trustor and the value received by the trustee are perceived equally. The curve with the accumulated gains can be seen in Fig. 3. The trustee marginally sees an increase in the received value, whereas the trustor sees a small decline.

This scenario could be acceptable depending on the scale of the service and number of trustees, since the trustor's final gain is the aggregate from all trustees.

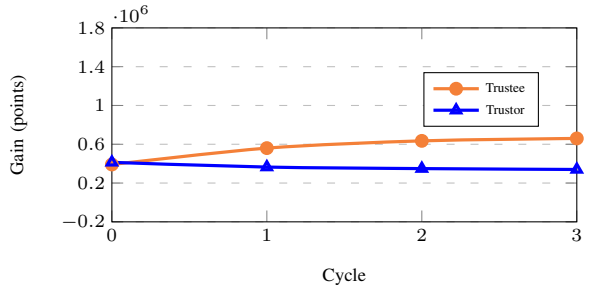


Figure 3. Accumulated Gains ($K = 1$).

C. Simulation 3: Machine Learning Inferences Are Causing Inefficiencies

The third simulation, Fig. 4, shows a scenario where inefficiencies are being brought upon the trustee ($0 \leq K < 1$). Their gains are at best negligible, and at the same time there is a significant drop in the trustor's gains.

This situation would be plausible and acceptable only during the development phase of a product, where a trustee would have accepted to be an early adopter of the service.

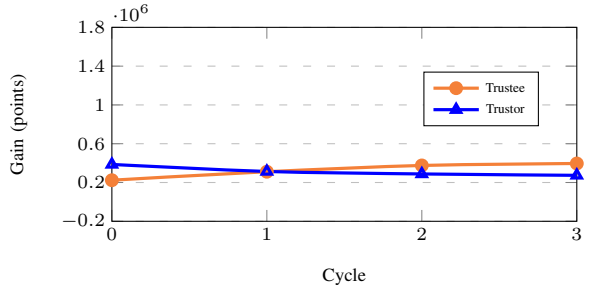


Figure 4. Accumulated Gains ($0 \leq K < 1$).

D. Simulation 4: Machine Learning Inferences Are Rapidly Eroding Trust

The last simulation shows the worst-case scenario where machine learning inferences erode the trustor's trustworthiness ($K < 0$), reducing the trustee's ability to trust. Fig. 5 shows how, in this scenario, there are negative gains (loss) for trustors and trustees. They are both worse off with the service, compared to operating without it.

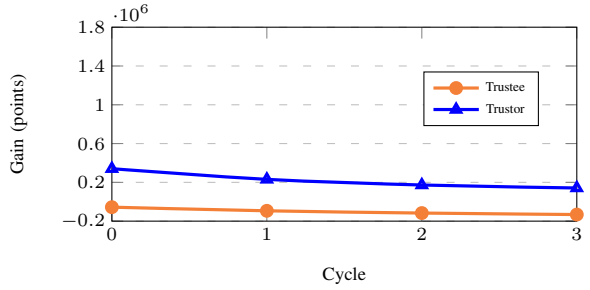


Figure 5. Accumulated Gains ($K < 0$).

VI. MEASURING AI AND ML RISK

One of the intended outcomes of quantifying trust is to define the metrics of risk. Then, it can be measured and monitored.

The National Institute of Standards and Technology (NIST) has published a study called "Artificial Intelligence Risk Management Framework (AI RMF)" [7]. There, they claim that there is a finite set of traits that approximate to a good definition for a system to be trustworthy. We aim to extend the concepts to implement quantitative metrics and create a viable framework to monitor trustworthiness. NIST identifies seven broad categories. They are (The color-coded categories will be useful later in this paper when understanding an example of the framework implementation):

1) Reliability and Validity
2) Safety
3) Security and Resilience
4) Accountability and Transparency
5) Explainability and Interpretability
6) Privacy
7) Bias Management

For each of those categories, this paper proposes metrics that can be measured and used to compute a *trust score*.

A. Reliability and Validity

A system is reliable when it does its job as intended, with minimal disruption of service [8], and when the results produced can be confirmed through objective evidence that the requirements were met [9]. The following are proposed metrics for reliability and validity:

- Uninterrupted uptime.
- Number of crashes.
- True Positives, True Negatives, False Positives, False Negatives.
- Latency between inquiry and returning results.
- Additionally, depending on the specific use case, the adoption of specific metrics (Accuracy, F1 [10], BLEU [11], SuperGLUE [12], HELM [13]) is encouraged.

B. Safety

The state of the data, the system, the people, and the subject of inferences are not at a meaningful risk, that extends beyond physical safety. Those are metrics to represent that:

- System design is represented in a diagram and is peer-reviewed, where appropriate.
- Data handling is done via a well-defined process with clear controls that align with existing regulations and oversight.
- A report that details to customers which data fields are used in training models.
- Access to the data is done with the consent of customers and is system-wide enforced by access roles.
- Once a model architecture is defined, models are trained using automation that does not require the intervention or participation of personnel.

C. Security and Resilience

Everyday operations have the ability to withstand adverse events or unexpected changes in the use or functioning of the environment.

- Systems and people have explicit credentials to run and/or access the data.
- Isolation of data and systems from unauthorized agents.
- Implementation of multiple scopes of granted access/runtime, with each agent being assigned the minimum necessary level to perform a task.

D. Accountability and Transparency

Accountability and transparency of operations are necessary conditions for being trustworthy and increasing trust.

- Report the data used in model training back to customers. The system must have a report, accessible by customers, that shows what data was used to train models.

E. Explainable and Interpretable

Explainability: the representation of the mechanisms underlying AI systems' operations [7].

Interpretability: the meaning of AI systems' outputs in the context of their designed functional purposes [7].

- Identification of the principal component of inference results.
- Displaying similar records can explain an inference by analogy.
- The ratio between the number of explanations given over the total number of explainable records would be the key metric.

F. Privacy

The norms and practices that help safeguard human autonomy, identity, and dignity. Freedom from intrusion, limiting observation, obtaining consent prior to disclosing or using Personally Identifiable Information (PII).

- Definition and implementation of Legal, Privacy, and Responsibility frameworks.
- Transforming raw data into embeddings.
- De-identification and aggregation.
- Privacy awareness and training of the people involved.

G. Bias Management

Establish reasonable and viable frameworks for error prevention, then optimization of execution for error correction.

- Number of reported and confirmed use cases.
- Subsequently, it can be offset by releasing new model versions that address those issues.

VII. IMPLEMENTATION

Each of the metrics discussed in the "MEASURING AI AND ML RISK" section become numeric entries in a vector M, and associated with it, there is a stochastic vector S containing weights representing how important each of the traits are in contributing to the creation of value and trustworthiness.

The dot product between M and S produces the *Trust Score* W (10) which is our metric to signify value and trust.

$$W = M \cdot S^{\mathsf{T}} \tag{10}$$

The following is a simulated example of numeric scores attributed to items in each of the 7 categories. The names next to the entries of vector M describe each item, as previously proposed, and the colors represent the categories.

The entries in vector M related to True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) share Number of Inferences as common denominator. It is important to clarify that this is the total number of inferences where it is possible to categorize them as [TP, TN, FP, FN]. In many cases, the categorization of an inference is unknowable.

Note that the stochastic vector S contains negative entries. Those are to penalize the corresponding metric in vector M and reduce the trust score. For example, the higher the number of crashes, the lower the score.

$$M = \begin{bmatrix} \text{Uptime} & 99.99\% \\ \text{Number of Crashes} & 3 \\ \text{True Positives/Number of Inferences} & 60.00\% \\ \text{True Negatives/Number of Inferences} & 34.29\% \\ \text{False Positives/Number of Inferences} & 4.29\% \\ \text{False Negatives/Number of Inferences} & 1.43\% \\ \text{System Design} & 1 \\ \text{Data Handling Processes} & 1 \\ \text{Data Points Report} & 1 \\ \text{Data Access Consent} & 1 \\ \text{Touchless Model Training} & 1 \\ \text{Access Control} & 1 \\ \text{Tiered Access} & 1 \\ \text{Data Isolation} & 1 \\ \text{Data Usage Report} & 1 \\ \text{Inference Explanation} & 40.00\% \\ \text{Present Similar Records} & 20.00\% \\ \text{Number of Explanation/Total Inferences} & 10.00\% \\ \text{Legal and Privacy Frameworks} & 1 \\ \text{De-identification of Data} & 0 \\ \text{Privacy Training} & 1 \\ \text{Number of Confirmed Bias Issues} & 2 \\ \text{Number of Deployed Bias Fixes} & 1 \end{bmatrix} \tag{11}$$

$$S = \begin{bmatrix} 0.14 \\ -0.14 \\ 0.24 \\ 0.24 \\ -0.10 \\ -0.10 \\ 0.01 \\ 0.02 \\ 0.02 \\ 0.01 \\ 0.04 \\ 0.06 \\ 0.07 \\ 0.06 \\ 0.05 \\ 0.05 \\ 0.05 \\ 0.05 \\ 0.06 \\ 0.05 \\ 0.02 \\ -0.06 \\ 0.06 \end{bmatrix} \tag{12}$$

Metrics of a qualitative nature are expressed numerically in vector M as 0 or 1, representing their absence or presence. For instance, in the "Accountability and Transparency" section, we mention implementing a report that shows what data was used in training models. Either the report is available, or it is not. Although there may be degrees of completion of adoption in an organization, we are focused on the customer's perspective, that either the item is in place or it is not.

If possible, it will be good to keep the trust score within the [–1, 1] range, where –1 is the worst possible score, and 1 is the best score. We can use (13) to apply this range constraint to the result of the *trust score* computation.

$$W = min(1, max(W, -1)) \tag{13}$$

In the case of the example provided in (11) and (12), the trust score would be:

$$W = 0.635557$$

A. Temporality

The trust score W is expected to display fluctuations over time. Since systems could experience an occasional malfunction, a model performance degradation, or an unanticipated incident, however, those fluctuations are presumed to be narrow and gentle, rather than wide and abrupt like a roller coaster.

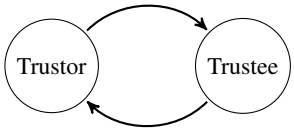
It is plausible to imagine that after a few cycles of significant fluctuations in the *trust score*, a customer would disengage and discontinue usage of the product.

VIII. FAIR TRADING

Fairness is an intrinsic concept associated with trust. Assuming that the trustor is providing value to a trustee, and in return the trustee is returning something of value to the trustor, the next step is to find that region of equilibrium where both parties accept the exchange as fair trade.

In addition, the region must be defined in such a way that it scales up or down proportionally to the exchange of value. For instance, imagine that a trustor went from providing one service, to providing two or three services; the trustor will expect to charge the trustee more. This section shows how this region of equilibrium is computed in such way that it remains a fair trade for both parties.

From (5) and (6), we know that the accumulated value A_u by the trustor is the product's residual value left, plus the repayment value sent by trustees. The net gain N_v by trustees is the received magnified value, minus the repayment.



$$A' = (1 - p)A + qN \tag{14}$$
$$N' = KpA - qN \tag{15}$$

From the previous paragraph, we see that (14) and (15) express how the next state of accumulation A' and net gain N' are computed. Expressing them in matrix format gives us (16).

$$\begin{pmatrix} A' \\ N' \end{pmatrix} = \begin{pmatrix} 1-p & q \\ Kp & -q \end{pmatrix} \begin{pmatrix} A \\ N \end{pmatrix} \tag{16}$$

We want to find that region of values that would make the trade between trustor and trustee to be considered fair.

From Linear Algebra, we know that the eigenvectors [14] of a matrix will give us the space that could scale—but otherwise would remain unchanged—irrespective of the linear transformation applied to it (assuming that the eigenvectors are linearly independent and have no imaginary i component).

Given that the conditions are satisfied, the linear transformation would be the addition or subtraction of services and the proportional increase or decrease of charges and feedback interactions, in other words, scaling up, down, or neutral.

The eigenvector associated with the largest, positive eigenvalue of the matrix shown in (16) can be interpreted as the region where both parties should consider transactions between them as fair trade, thus contributing to preventing the erosion of trust.

Let us build an example. Assume that the percentage of remitted value p, the repayment portion q, and magnification factor K have the following values:

$$p = 0.85,\ q = 0.14,\ K = 2$$

Substituting these values in the matrix from (16) leads us to:

$$\begin{pmatrix} A' \\ N' \end{pmatrix} = \begin{pmatrix} 0.15 & 0.14 \\ 1.7 & -0.14 \end{pmatrix} \begin{pmatrix} A \\ N \end{pmatrix} \tag{17}$$

One condition that needs to be satisfied is that the vectors—derived from the matrix in (17)—are linearly independent, so they can span the space being considered. Otherwise, they would only represent a sub-space and not necessarily produce the fair trade region we aim for.

As you can see in Fig. 6, the vectors are linearly independent and also satisfy the other conditions to compute the eigenvectors to determine the fair trade region between trustor and trustee.

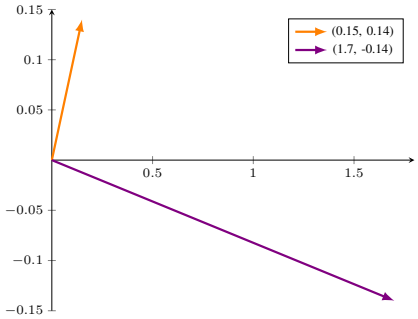


Figure 6. Linearly independent vectors.

The next step is to compute its eigenvalues and eigenvectors, then find the line defined by the eigenvector associated with the largest eigenvalue.

$$\lambda_1^{\downarrow} = 0.513945 \tag{18}$$
$$\lambda_2 = -0.503945 \tag{19}$$
$$E_{\lambda_1} = \begin{pmatrix} 0.384674 \\ 1 \end{pmatrix} \tag{20}$$
$$E_{\lambda_2} = \begin{pmatrix} -0.214085 \\ 1 \end{pmatrix} \tag{21}$$

The largest eigenvalue is $\lambda_1^{\downarrow}$, thus our eigenvector of interest is E_{λ_1}. In order to find the line defined by E_{λ_1} coordinates, we just need to compute its slope, since the eigenvector starts at the origin (0, 0).

$$y = mx + b \tag{22}$$

$$m = \frac{1-0}{0.384674-0} = 2.599604 \tag{23}$$

$$b = 0 \tag{24}$$

$$y = 2.599604\,x \tag{25}$$

Fig. 7 shows the eigenvector E_{λ_1} in red and the line derived by it, and defined by (25), in blue. The line characterizes the fair trade region since any point on it carries the maximum accumulated value A and net gains N, for the trustor and trustee, respectively.

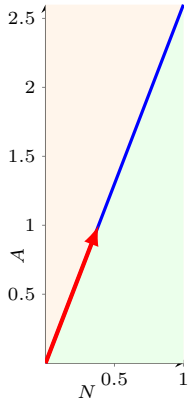


Figure 7. Fair trade region.

The colored areas above and below the line represent the regions where either the trustor would accumulate more value (orange) or the trustee would retain more gains (green).

IX. Conclusion

This paper takes a step forward in contributing to the conversation about trust in ML-based systems. It presented a realistic and viable framework to compute a trust score and demonstrated that good machine learning inference results satisfy a valid criterion to increase a trustor's trustworthiness, allowing for trustees to be more trusting.

A strong motivation exists to provide inferences only when a minimum confidence level has been cleared. It would be preferable to not produce a result than to provide a low-confidence one. When nothing is provided, a customer can still operate at their nominal level of productivity.

We established the items of interest for measuring, defined a system to compute and weigh each contribution, and identified the region of fair trade where win-win relationships between trustor and trustee can take place and scale up or down.

Trust has a temporal nature to it; its behavior is not linear, but instead it is expected to oscillate with gentle fluctuations. Trust and value add are not only earned, but also require maintenance over time.

Lastly, we demonstrated that it is possible to establish a region of fair trading where both trustors and trustees perceive fairness in the exchange of value.

References

[1] K. Martineau, "What is AI inferencing?" 2023. [Online]. Available: https://research.ibm.com/blog/AI-inference-explained (Last accessed: 2024-05-14)

[2] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity, and social history," *Games and Economic Behavior*, vol. 10, no. 1, pp. 122–142, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0899825685710275 (Last accessed: 2024-05-14)

[3] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[4] N. Gillespie, S. Lockey, C. Curtis, J. Pool, and A. Ali, "Trust in artificial intelligence: A global study," KPMG - The University of Queensland, Tech. Rep., 2023. [Online]. Available: https://doi.org/10.14264/00d3c94 (Last accessed: 2024-05-14)

[5] D. Kreps, "Corporate culture and economic theory," *Perspectives on Positive Political Economy*, pp. 90–142, 1990.

[6] D. Cirne. (2023, 09) Simulations source code. [Online]. Available: https://gist.github.com/dcirne/8c74a2d8d5adaf59f9366a5212d41f22 (Last accessed: 2024-05-14)

[7] NIST, "Artificial intelligence risk management framework (ai rmf 1.0)," Tech. Rep., 01 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (Last accessed: 2024-05-14)

[8] "Trustworthiness — vocabulary," International Organization for Standardization, Geneva, Switzerland, Technical Specification ISO/IEC TS 5723:2022(en), 2022.

[9] "Quality standard," International Organization for Standardization, Geneva, Switzerland, Technical Specification ISO/IEC TS 9001:2015(en), 2015.

[10] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," 2014. [Online]. Available: https://arxiv.org/pdf/1402.1892.pdf (Last accessed: 2024-05-14)

[11] K. Papineni, S. Roukos, T. Ward, and W. Zhu "BLEU: a Method for Automatic Evaluation of Machine Translation." IBM, 2022. [Online]. Available: https://aclanthology.org/P02-1040/ (Last accessed: 2024-06-17)

[12] A. Wang et al, "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems", 2019. [Online]. Available: https://w4ngatang.github.io/static/papers/superglue.pdf (Last accessed: 2024-06-17)

[13] P. Liang et al, "Holistic evaluation of language models," 2022. [Online]. Available: https://arxiv.org/pdf/2211.09110.pdf (Last accessed: 2024-05-14)

[14] G. Strang, *Linear Algebra and Its Applications*, 4th ed. Cengage Learning, 2006.

[15] E. Parliament and C. of the European Union, "Regulation (EU) 2023/656 of the European Parliament and of the council of 14 June 2023 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act)," pp. 1–231, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R0656 (Last accessed: 2024-05-14)

[16] C. Alós-Ferrer and F. Farolfi, "Trust games and beyond," *Frontiers in Neuroscience*, vol. 13, 09 2019. [Online]. Available: https://doi.org/10.3389/fnins.2019.00887 (Last accessed: 2024-05-14)