Fostering Trust on Machine Learning Inferences

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

***Abstract*—Artificial Intelligence (AI) and Machine Learning (ML) providers have a tremendous responsibility to develop valid and reliable systems. Much is discussed about trusting AI and ML inferences, but little has been done to define what that means. Those who work in the space of ML-based products, have familiarity with the topics of transparency, explainability, safety, bias, and so forth, yet there are no frameworks to quantify and measure such items. 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 further 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. Then, we conclude by proposing a framework to capture quantitative metrics to be used to objectively understand how trustworthy an AI and ML system can claim to be, and their trend over time.**

***Index Terms*—artificial intelligence, game theory, machine learning, trust.**

I. Introduction

The emergence of a new technology has always been accompanied by the challenge of earning the trust of the public in general. This has happened during the Industrial Revolution—with the mechanization of processes—and in numerous other occasions. Today, our problem is to establish a framework for increasing trust on systems powered by machine learning.

Much talk has taken place, but not much has been done to establishing a framework to measure trust.

In this paper, a step is taken in the direction of defining a framework for quantifying and measuring trust on machine learning inferences. This, however, has its own challenges, such as: defining a starting point, measuring qualitative aspects, and tracking the trust level over time.

The paradigm explored here assumes that trust is built by an initial *altruistic* act by the trustor, signaling that the actor is trustworthy. More specifically, the trustor's altruistic 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.

The objective is to make customers trust—above a minimum threshold T—as to incentivize them to engage in the *Trust Game* [1]. These games are extensions built on top of the foundation of *Game Theory* [2].

In addition, 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 the interactions, where both actors collaborate and observe each other, reacting to historical actions from one another.

Models are representations that aspire to approximate reality, and like other models, the framework proposed here is subject to the noise in its variables, and the gap between what is captured versus what actually happens. The fewer distortions, the better the framework becomes.

In the *Trust Games* section, we establish the flow of how the value of a product is transferred to trustees, and how trustors receive a portion of that value back. Then, we propose a numeric framework to measure the trust level in the *Quantifying Trust* section. Next, we define the criteria for obtaining a minimum level of trust in the *Threshold* section, and last, in the *Simulated Experiments* section, we conclude by presenting the results from the simulations [3].

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.

Note that the nature of concurrency allows for independent actors (trustees) to observe and react to the actions of other actors.

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:

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

ISBN: 978-1-68558-089-6

Depending on the quality of the results delivered by the trustor, the perception of value by trustees 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. They may decline continuing the trial or decline a contract 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; 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.

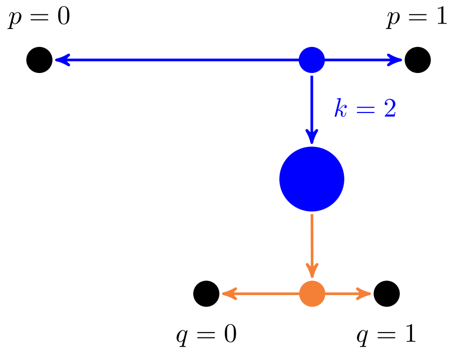

Fig. 1: Trust Game payoffs.

Regarding the magnification factor, for the cases where $K > 1$, 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 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 continuous and repeated interactions between trustor and trustee that take place over long periods of time, with no specified temporal upper boundary.

After the initial remittance R_u (1), there may be 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:

$$\begin{aligned} N_v^{\text{1st cycle}} &= G_v^1 - q_1G_v^1 \\ &= (1 - q_1)K_1p_1V \end{aligned} \tag{6}$$

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

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

and trustee:

ISBN: 978-1-68558-089-6

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

The objective is to maximize the payoff to trustee and trustor, establishing a region where the exchange of values is considered fair trade. As such, trust must be repaid [4] (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:

$$W_u \subseteq \begin{cases} pV \geq T \\ K \geq 1 \end{cases} \tag{9}$$

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 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. Simulated Experiments

The following are a set of four experiments that simulate scenarios from fostering to eroding trust as a result of the quality of machine learning inference.

All the experiments 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: forty-two, nine thousand, or seventy-three billion. What we want to observe is the shape of the curve formed from plotting interaction cycle after interaction cycle.

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

In each of the experiments, we observe the shape of the curves and their accumulated trend iteration after iteration. Also, 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 experiment, we will go step-by-step in the first interaction. For subsequent experiments, only the final graph plots will be shown. Irrespective of the experiment, they all can be reproduced using the source code [3] that accompanies this paper.

Here, the assumption is that machine learning inferences are magnifying the value of the product ($K > 1$).

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$1,000,000), the trustor's accumulated gain is equal to 532,000 ($A_u =$ 350,000 + 182,000) points. And the trustee's gain is 1,118,000 ($N_v = 0.86\times$ 1,300,000) points.

First, the trustee's perception was that they received more value that 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). 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 experiment.

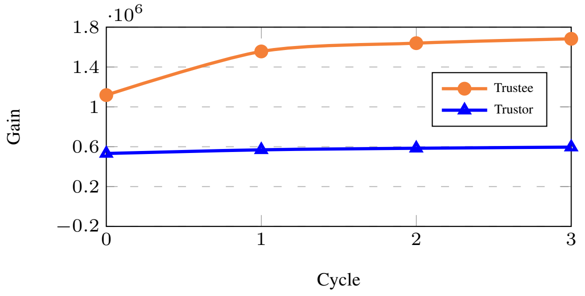

Fig. 2: Accumulated gains ($K > 1$).

B. Simulation 2: Machine Learning Inferences Are Neutral

For the second experiment, 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.

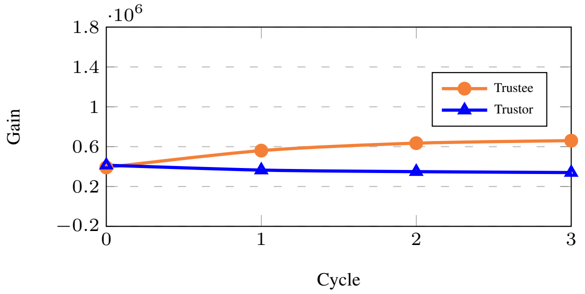

Fig. 3: Accumulated Gains ($K = 1$).

C. Simulation 3: Machine Learning Inferences Are Causing Inefficiencies

The third experiment has a curve (Fig. 4) showing 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. Otherwise, there would be no return on investment to the trustee, and a loss of value to the trustor.

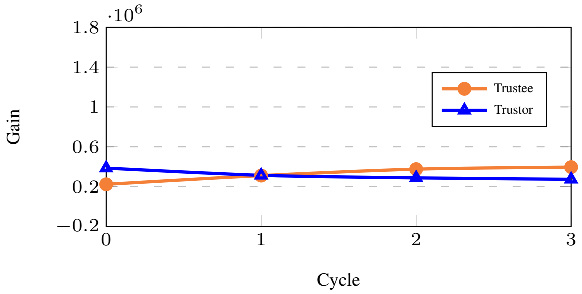

Fig. 4: Accumulated Gains ($0 \leq K < 1$).

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

The last experiment shows the worst-case scenario where machine learning inferences are eroding the trustor's trustworthiness ($K < 0$), therefore reducing the trustee's ability to be trusting. Fig. 5 show 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.

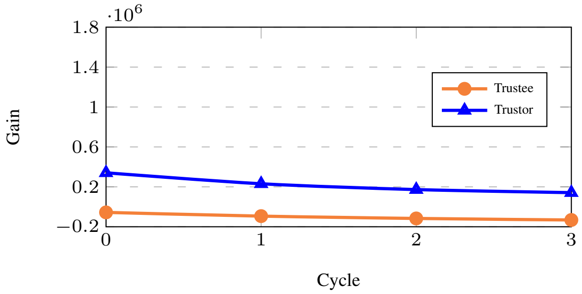

Fig. 5: Accumulated Gains ($K < 0$).

VI. Conclusion

This paper takes a step forward in contributing to the conversation to define, in a quantifiable manner, what trust in ML-based systems means. Here, we demonstrated that good machine learning inference results satisfy a valid criterion to increase a trustor's trustworthiness, allowing for trustees to be more trusting.

There exists a strong motivation for ML-based products 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.

Plans for future research include: 1- proposing a set of criteria to define the risk associated with an inference; 2- establish a quantitative process to measure it.

References

[1] 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

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

[3] D. Cirne. Simulated experiments source code. [Online]. Available: https://gist.github.com/dcirne/8c74a2d8d5adaf59f9366a5212d41f22

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

[5] C. Alós-Ferrer and F. Farolfi, "Trust games and beyond," *Frontiers in Neuroscience*, vol. 13, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2019.00887

ISBN: 978-1-68558-089-6