The Cycle of Trust and Responsibility in Outsourced AI

Maximilian Castelli and **Linda C. Moreau, Ph.D.**
Amazon
Herndon, VA, USA
(maxcaste, lcmoreau)@amazon.com

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) are rapidly becoming must-have capabilities. According to a 2019 Forbes Insights Report, "seventy-nine percent [of executives] agree that AI is already having a transformational impact on workflows and tools for knowledge workers, but only 5% of executives consider their companies to be industry-leading in terms of taking advantage of AI-powered processes." (Forbes 2019) A major reason for this may be a shortage of on-staff expertise in AI/ML. This paper explores the intertwined issues of trust, adoption, training, and ethics of outsourcing AI development to a third party. We describe our experiences as a provider of outsourced natural language processing (NLP). We discuss how trust and accountability co-evolve as solutions mature from proof-of-concept to production-ready.

1 Introduction

Our business unit specializes in providing AI/ML solutions to customers seeking to use NLP and other AI capabilities to augment human analysts. Our typical use case involves customers with a small number of highly specialized subject matter experts (SMEs) who need to assess a large number of documents in a short amount of time, often in the context of high-stakes missions. Our third-party NLP solution space is comprised of secure, cloud-deployed processing pipelines that transform unstructured text collections into actionable insights using combinations of customized entity extraction and text classification. The pipelines produce a sortable and filterable data stream suitable for prioritized review by analytic end-users.

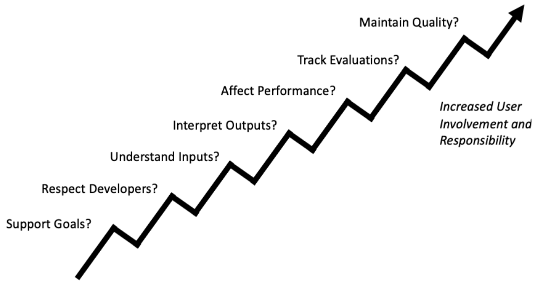

Figure 1 Trust Growth in the AI Adoption Journey

With increasing frequency, we are approached by customers who have heard of our early successes in AI-based analyst augmentation and would like to achieve similar results in their own operations. Regardless of perceived similarities among new opportunities and previous successful applications, we believe that ethically, the responsibility lies with us to assure that each potential use case for AI/ML adoption is appropriate, feasible, and sustainable. Feasibility and sustainability play into the ethics of AI/ML solutions and also in our ethical dealings with customers. This includes ensuring that our customers have appropriately managed expectations for what AI can and should do in their context. It means working to understand the specifics of customers' requirements, including:

- the nature of their data
- the questions they need to ask of the data
- the availability of legacy data usable for training and evaluation
- the availability of SMEs to validate models
- the potential development of feedback loops for continuous model improvement.

Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), pages 43 - 48
July 14, 2022 ©2022 Association for Computational Linguistics

For the outsourced development case, responsible engagement also involves assessing the ability of the customer's staff to perform a few critical types of functions once their engagement with the outsourced team has ended: 1. They must be able to offer training to their end users about how to properly and ethically interpret model outputs; 2. They must understand that unintended consequences may arise if they try to retarget a model trained on one type of input for use on some other type of data, and, 3; . They must be able to operate and maintain the NLP pipeline and its models. The latter requirement includes many sub-tasks, including the ability to: react to deficiencies in the model; detect model drift; manage model versioning; and retrain models as necessary. All of these questions relate to the customer's AI literacy. If the AI literacy of the receiving team is low, it would seem that the delivery team has a greater ethical responsibility for providing education, guidance and possibly ongoing support.

We explore anecdotes from one of our earliest engagements to highlight various facets of trust, ethics, and, responsibility that have arisen via our experience as a third-party provider of AI/ML-based NLP. Though we use this initial engagement as a backdrop for our discussion, we have observed this pattern repeatedly across a variety of subsequent customer engagements. Based on this cumulative experience we have begun preparing an "AI/ML adoption framework" consisting of various knowledge elicitation artifacts, including questions to pose at different stages of development. These can help ensure consistent and responsible assessment of the AI-readiness of new and existing customers. We posit that trust in AI-based solutions can be effectively built through a cycle of engagement among the AI solution providers, the end users, and the models. Users can cyclically build ownership, accountability and the understanding required for explainability by being actively engaged in model-building and maintenance. We also point out critical questions that must be posed throughout the development lifecycle to maximize adoption of AI and the infrastructure in which it is embedded.

2 An Eye Opening First Engagement

One of our earliest customer engagements corresponded closely to the typical use case described in the introduction, in which we augment human workflows with automated NLP processing using a framework like that depicted in Figure 2. This particular engagement was small in terms of data size, typically under 10k documents per batch, but was nonetheless extremely impactful. The work was initiated by decision makers who believed that an AI-based solution using cloud services would provide a much-needed productivity boost for their highly valuable, specialized, yet under-staffed analytic workforce. They engaged our team to help create a cloud-based data processing pipeline to automate some analytic tasks performed by their staff, hoping to free up the SMEs to focus on other less automatable duties.

Our first trust-building challenge arose from our customer's initial expectation that our analytic pipeline would be fully automated, removing the human analyst from the loop completely. It became rapidly apparent that full task automation would not be advisable any time in the foreseeable future. The existing body of labeled data was produced by a single analyst responsible for producing a binary classification indicating whether or not reports were relevant to the team's mission. On a monthly basis, the analyst's process was kicked off by manual execution of a standing Boolean database query. The Boolean query returned an unranked list of documents numbering in the thousands or tens of thousands. Each document would then be manually reviewed and tracked in a spreadsheet. Any report deemed of interest would be subsequently annotated to highlight entities and key phrases of interest. The analyst would then generate visualizations to communicate his findings. Processing a typical tranche of data in this way would take that SME analyst a minimum of three full business days, but often much more for larger document sets.

This background scenario meant that we were starting our ML development with an initial data set that had an extremely skewed distribution of relevant reports to non-relevant ones. The data also featured annotations that had been produced without the benefits of standard annotation guidelines and automation. Given this, we needed to help our customers understand why it would be beneficial to opt for a user-in-the-loop, active learning scenario. Fortunately, the decision makers and the analyst grasped the proposal immediately and were eager to help create a sustainable solution.

2.1 Early-stage Trust Building

The need to assure that AI-based NLP solutions are appropriate for a customer, and to help that customer and all of their stakeholders develop trust in the solutions has become a recurrent theme in our engagements. The customer teams are never monolithic, meaning that although some members may embrace an AI-automated solution from the outset, others are leery of any modification to their workflows and of any suggestion that a machine can "do their job". The onus is on the third-party delivery team to address concerns of AI/ML applicability, and to help raise AI literacy of all of the stakeholders. This means not only being mindful of the existing team and their traditional workflows, but also being careful to explain and demonstrate the new capabilities in context. For our initial engagement, a key element of trust building involved frequent consultation with the customer SME team to make sure their workflow preferences were respected and that data annotation requirements were accurately captured and codified.

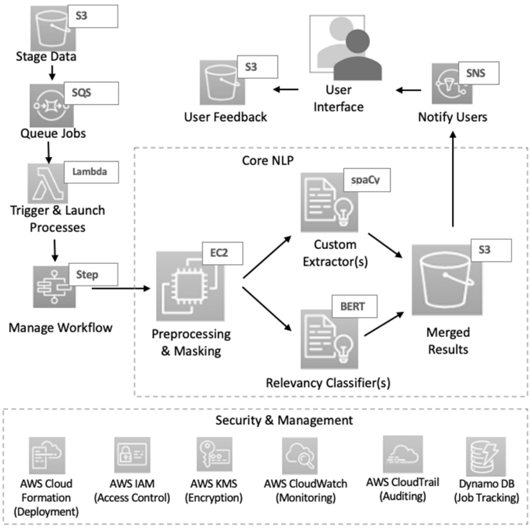

Figure 2 NLP Pipeline Data Flow

One example of evolving user expectations and building trust arose from experimenting with several alternative preprocessing techniques. The customer team had a preconceived notion that the relevancy classification models would perform better with the inclusion of all available document metadata, including numerous repetitive and verbose attributes. We addressed this assumption by doing a detailed breakdown of how different preprocessing techniques affected model performance, presenting iterations of confusion matrices and file outputs based on the differing levels of document preparation. This exercise in providing transparency through evidence was really a first step in gaining the SME's trust, and it laid the groundwork for collaboration. Rather than dictating to the end users what needed to be done, we took the time to bring them along in their understanding. This led to a cycle of knowledge elicitation and feedback in which our SME user-base understood the development process and took ownership of the quality of the pipeline outputs. With their collaboration, we developed a pipeline designed to augment their workflow, yet keep them empowered and in control of their data. They were able to benefit from a host of data enrichments and model inferences. We find these anecdotes to be powerful examples of techniques for early-stage trust development. The main idea is to increase the customer's AI literacy over time, providing them understanding of the data and their critical role in enhancing it.

2.2 The Product – The NLP Pipeline

When evaluating the ethical delivery of a third-party solution, the equation must weight impact and adoption equally – does the solution accelerate the customer's business, and can they successfully use it in their day-to-day work? For this engagement, we built a data processing pipeline using AWS cloud infrastructure and many of its data security features, as depicted in Figure 2. The overarching design principle was to use as much serverless workflow computing as possible, since this is generally less costly for the customer than running full virtual machines. The machine learning components, which require more compute power, were run on EC2 instances. Custom classifiers and named entity recognizers based on BERT (Devlin et al. 2019) and spaCy (Honnibal et al. 2021), respectively, were run on EC2 servers appropriately sized to meet their processing requirements. For building user trust in the pipeline, visibility into the data and model performance was provided by the customer-facing user interface (UI). From this UI the SME users could perform all of their normal job functions, access NLP pipeline outputs, and add ground truth labels all from a unified, familiar UI. This is important because it offered the most minimally-invasive augmentation of their existing workflows, meaning they never felt that the AI "got in their way". This is an important lesson learned – the NLP tools should remain as unobtrusive and easy-to-use as possible to avoid slowing down their acceptance. On the flip side of that, the increased efficiencies gained by the introduction of document ranking, classification and term highlighting played role in the acceleration of their trust in the NLP solution. This user acceptance ensured a successful path to continuous machine learning.

2.3 Continuous Improvement Builds Trust

It is imperative that AI-based systems continue to evolve and improve over time, or their value and user confidence will wane. For practitioners of AI/ML, the benefits of continuous learning may seem obvious. For our end users in this engagement, the benefits came as a highly motivating, pleasant surprise. Thanks to the unintrusive UI for capturing SME ground truth, the team was able to quickly produce demonstrable progress, significantly boosting NER model performance over a few months from an initial F2-SCORE of 0.51 to a more acceptable 0.87 for NER. A similar pattern occurred with the BERT relevancy classification models, where, through a combination of enhanced preprocessing and ground truth augmentation, we were able to boost model performance from 0.73 to 0.91F2. These highly visible improvements, which produced outcomes increasingly aligned with analyst intuitions, motivated the SME team to continue providing model feedback, despite the addition of some additional steps to their daily workflow. As Alon et al. (2020) state, a model is more trustworthy when the observable decision process of the model matches user priors on what this process should be. Thus, showing performant metrics on both historical and emergent data goes a long way toward cultivating trust in the pipeline and its models.

This discussion has highlighted the importance for trust building of demonstrating continuous improvements to the user, and of helping the user understand NLP and their role in improving it. Based on lessons learned from the first engagement, we now insist on the routine incorporation of AI-literacy materials and tools as part of our deliverables, including such artifacts as runbooks, annotation guidelines, and robust documentation to enable ongoing customization and enhancement of models by inheriting teams.

2.3.1 Operationalization

So far, we have focused on how we built trust in the NLP capabilities of a specific early engagement, and described how we have begun applying our lessons learned to subsequent engagements. One very important measure of the successful adoption of the initial system, and by extension of trust in NLP, was that it has led to four, and counting, additional applications using the same pipeline architecture pictured in Figure 2. The new uses cases, of course, have NLP components (spaCy and BERT) that are custom tailored for additional missions and end users. Despite starting from a higher initial level of trust thanks to the first success story, each additional use case has required a novel cycle of trust building and user adoption.

The positive impacts of putting the first NLP pipeline system into operational use were many. For the decision makers who commissioned the work and for the end users, the most obvious impact was in speed. Their time to process decreased from a minimum of several days to under a few hours for tens of thousands of documents. This speed-up, coupled with the document prioritization based on AI/ML-based inference results, led to multiple high value findings being brought forth quickly, within an impactful, actionable period of time.

A less obvious but equally valuable outcome of this operationalization lay in the knowledge capture implicit in the active learning cycle. Previous to the deployment of the system, the SME insights and intuition were only indirectly captured for positive exemplars in the form of unstructured analytic reports. The feedback loop in the pipeline now captures labels for both positive and negative examples and collects annotations for the named entities and key phrases that signal mission relevance for the SME.

Based on lessons learned from the initial success story and the follow-on use cases, we maintain the important principle of designing operational systems that incorporate continuous ML into the end user's existing workflow in as unobtrusive yet transparent a way as possible. Offering model transparency has meant experimenting with techniques for revealing clues about how the pipeline inferences have been achieved. This includes highlighting extracted entities, and also demonstrating on-demand visualizations using tools such as the Language

Interpretability Tool (LIT) (Tenney et al. 2020) and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016). We observed that LIT, LIME and similar interpretability assistants go a long way toward demystifying the "black box" for the end users and instilling them with confidence that there is human interpretable evidence available to support their further human analysis.

2.4 Transfer of Responsibility

Up to this point, we have focused on successful adoption of AI by decision makers and end users, and on our ability to build trust in NLP systems through introspection, transparency and user engagement. We now turn to more subtle ethical questions that arise under our business model of outsourced AI development. There are, of course, the normal software challenges of designing in a modular way to facilitate swapping of models and engines. Similarly, all such knowledge transfer requires thorough and well-written documentation. Beyond these usual concerns, though, are AI-specific considerations.

Customers receiving AI solutions need to understand various facets of ML operations. They need to be aware of the risks of model drift and understand the potential sources and impacts of model bias. They should be prepared to detect and mitigate those impacts. They need to be equipped with the knowledge and tools required to implement best practices in model management, including meticulous tracking of model inputs, processing procedures and parameters. They need evaluation infrastructure and an understanding of how to manage ground truth data.

How to best address these concerns remains an open question. It is essentially asking customers to either hire new staff with the appropriate expertise or to train their existing staff to become experts in machine learning. Another possible approach is for AI delivery teams to offer Operations and Maintenance (O&M) services on retainer to guarantee that the systems we create continue to operate to the highest possible standards. There is a blurry line of responsibility between those who commission AI systems and those who create them. Both parties must work together to achieve model sustainability and ethical usage. Underpinning this collaboration must be direct communication about the nature of the challenges.

We have attempted to address these concerns using a combination of architecture, documentation and education. Similar to the findings of (Srinivasan & de Boer 2020) regarding auditability, we placed extra emphasis on auditing all changes and assumptions made with the data and models in order to build customer trust in the solution and development processes. We have also built in a knowledge-transfer phase at the end of every engagement, which intersperses technical exchanges, Q&A sessions, and guided, hands-on use by the receiving team of tooling for model retraining and deployment. Ethical transfer of statistical models in these scenarios requires commitment to knowledge transfer and education.

3 Conclusion

Our goal has been to highlight important questions of trust, ethics, and responsibility that have arisen via our experience as third-party providers of AI/ML-based NLP. We have discussed how user engagement and accountability co-evolve with trust as a capability matures from proof-of-concept to production-ready. We conclude by listing a few of the key questions to be posed at various phases of a responsible engagement.

- Is an AI/ML solution appropriate to the customer's use case?
- What is the technical depth of stakeholder team and how can we architect a solution they can both use and maintain?
- How can we teach end users to ethically interpret and employ model outputs?
- What combination of workflow and tools will help earn trust in the AI?
- What is our responsibility for assuring that the inheriting team can obtain technical resources for O&M of ML models?
- How much time should we reserve for knowledge transfer to ensure continued success with CI/CD best practices?

By attending to these types of questions from the outset of each engagement and throughout, we strive to maximize successful NLP deployment and to build long term trust in AI/ML and NLP.

References

Alon Jacovi, Ana Marasović, Tim Miller, Yoav Goldberg: "Formalizing Trust in Artificial

Intelligence: Prerequisites, Causes and Goals of Human Trust in AI", 2020; arXiv:2010.07487.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

Forbes Insights (2019). "Everyday AI: Harnessing Artificial Intelligence to Empower the Knowledge Worker" downloaded on 04 April 2022 from http://info.microsoft.com/rs/157-GQE-382/images/EN-CNTNT-Whitepaper-HarnessingAItoEmpowertheKnowledgeWorker.pdf.

Matthew Honnibal and Ines Montani. "Release v3.0.0: Transformer-based pipelines, new training system, project templates, custom models, improved component API, type hints & lots more · explosion/spaCy". GitHub. *Retrieved 2021-02-02.*

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, Ann Yuan: "The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models", 2020; arXiv:2008.05122.

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin: ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", 2016; arXiv:1602.04938

Vasan Srinivasan A, de Boer M (2020) Improving trust in data and algorithms in the medium of AI. Maandblad Voor Accountancy en Bedrijfseconomie 94(3/4): 147-160. https://doi.org/10.5117/mab.94.49425