Working with Beliefs: AI Transparency in the Enterprise

Ajay Chander
Fujitsu Laboratories of America
Sunnyvale, California, USA

Ramya Srinivasan
Fujitsu Laboratories of America
Sunnyvale, California, USA

Suhas Chelian
Fujitsu Laboratories of America
Sunnyvale, California, USA

Jun Wang
Fujitsu Laboratories of America
Sunnyvale, California, USA

Kanji Uchino
Fujitsu Laboratories of America
Sunnyvale, California, USA

ABSTRACT
Enterprises are increasingly recognizing that they must integrate AI into all of their operational workflows to remain competitive. As enterprises consider competing AIs to support a particular business function, explainability is an advantage which gets a candidate AI a foot in the door. Our experience working with enterprise decision makers considering AI in a decision augmentation role reveals an additional and possibly more crucial aspect of choosing an AI: the ability of decision makers to *interact fluidly* with an AI. Fluid interactions are necessary when an AI's recommendation does not match a human decision maker's existing beliefs. Interactions that allow the (typically non-technical) human to *edit* the AI, as well as allow the AI to *guide* the human, enable a collaborative exploration of the data that leads to common ground where both the AI and the human beliefs have been updated. We outline an illustrative example from our experience that models this dance. Based on our experiences, we suggest requirements for AI systems that would greatly facilitate their adoption in the enterprise.

Author Keywords
AI; Transparent AI; Accessible AI; Explainable AI; Interactive AI; Tunable AI; Beliefs; Enterprise

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; I.2.m. Artificial Intelligence: Miscellaneous.

INTRODUCTION
We offer this position paper to the community to share our observations and learnings from our vantage point of being the R&D arm of a global "top 5" IT behemoth. Our parent company is active across a very wide spectrum of IT products, technologies, and services, and in that role interacts with a large variety of enterprises globally. As improvements in the abilities of AI systems, in particular the dramatic improvements in the performance of machine learning systems have captured the popular imagination, enterprises worldwide have accepted the premise that an *Augmented Intelligence* enterprise is a necessity to survive and compete in the modern digital era. An enterprise with augmented intelligence, wherever possible:

1. Augments human sensing with sensors (IoT)
2. Augments human decision making with AI, and
3. Augments human action with software and hardware robots.

The process of onboarding an enterprise to augment its human decision making with AI typically follows a predictable script. A very common set of questions is asked by enterprise clients, typically comprising of:

1. What can AI do for me and for my enterprise?
2. Is there an AI system that can improve aspect X of my enterprise's workflow Y?
3. How do I choose, personalize, and integrate the system in (2) above into my enterprise?

The answer to the first question – presented through the capabilities of AI systems on external datasets – broadens awareness at the highest levels of the typical enterprise to the possibilities of modern AI, especially modern machine learning (ML) systems. This typically leads to the second question, which brings focus to a particular workflow Y in the enterprise. When presented with a few candidate AI systems that can improve this workflow Y, *some explainability* of the AI system – typically around a pre-selected dataset and prediction use-case – is table stakes today. This builds some assurance in the client that they are not bringing into their enterprise a runaway digital decision maker. The final step involves a detailed evaluation of the candidate AI system(s) using data proprietary to the enterprise, which is typically handed off to the corresponding leadership team and the human decision makers within it.

It is the *perceived* capabilities of the AI system in this third step that determine its eventual adoption in the enterprise. The stakeholders evaluating the AI in this stage are typically business domain experts but generally not technical experts. They generally have some strongly held *business beliefs* about their domain, for example, about how to approach sales in a particular region. These beliefs are borne out of their collective professional experience, and sometimes

obtained at significant economic cost. Hence, they tend to be sticky. A candidate AI may make a recommendation that is aligned with or aligned against the belief of the business stakeholder. When the AI is aligned with the business stakeholder, it may be reviewed weakly and its institutionalization may further existing biases as reflected in the datasets. In this case, it is especially valuable for the AI system to include bias determination [11, 12] so that they may alert around biased beliefs. When the AI is aligned *against* the business stakeholder, it tends to receive special scrutiny. In this case, it is crucially important that the business stakeholders (i.e., the human decision makers) can interact with the AI fluidly as they would with an external human consultant who gives them news that they may not like at first. In both cases, a successful AI system in the enterprise is a *Belief Worker*: it has to learn and stay aware of institutional beliefs, and assist in updating them by being accessible to a wide variety of potential enterprise users that may come to rely on it.

TRANSPARENT AI

In our experience, the practical adoption of AI systems in enterprises that are making the move to Augmented Intelligence depends on empowering not just AI engineers but crucially System Integration (SI) engineers and business stakeholders. Current AI systems, which involve primarily an AI engineer as the "human-in-the-loop", leave out these important constituencies. Based on our experiences, we posit the follow 4 pillars of Transparent AI:

1. *Accessible AI.* SI engineers and business stakeholders should be able to *ask* questions of AI without going through the AI engineer's interface. Progress in this area is most robustly being led [2] by the industry, because there is commercial demand for this.
2. *Explainable AI.* The *answer* that the AI comes back with should be accompanied with some explanation, as the audience for this answer is now no longer just the AI engineer. Progress is this area is most robustly being led by DARPA's XAI [1] project.
3. *Interactive AI.* The non-AI engineer does not have a dataset to evaluate the AI's answer against. What they do have is beliefs. It should be possible for the non-AI engineer to interact fluidly with the AI system to edit the AI, perhaps by editing its dataset in response to its answers. This process would continue until either the AI is updated or the beliefs are updated or both.
4. *Tunable AI.* Interactive AI systems enable a motivated user to update an AI through easy interactions. Taking that a step further, Tunable AI refers to sets of technologies that can, given an AI system, automatically identify usable "tuners" for an AI that can be utilized by *end-users*.

We call these sets of technologies collectively Transparent AI. The rest of our paper will describe aspects 1-3 of Transparency in the context of an example using a platform that we built called AI.AI, short for Accessible and Interactive AI.

RELATED WORK

DARPA's XAI initiative [1] has ignited broad interest in exploring issues related to the transparency of AI models. As AI is increasingly integrated into a wide variety of settings, from enterprise assistants to self-driving cars, a wide variety of users are now interested in understanding the decisions of AI systems. Accordingly, various notions of transparency are emerging across different application domains and different end-user types. A summary of the feasibility and desirability of transparency related notions from an AI engineer's perspective is offered in [3]. In [5], the authors propose a general taxonomy for the rigorous evaluation of interpretable machine learning. A survey of the desired features of transparent AI systems as viewed from a social and behavioral sciences perspective is provided in [4]. Below, we organize other related work within the 4 pillars of Transparent AI.

Accessible AI: Amazon recently announced the release of a service called Sagemaker [6], a framework for developers and data scientists that helps manage the systems infrastructure involved in starting and running AI pipelines. DataRobot [2] offers an automated machine learning platform as well as services and education to jumpstart AI related processes. There are many more such services in the offing.

Explainable AI: The usefulness of explainable models has been demonstrated across various application domains such as recommendation systems [16] and healthcare [17], to just name a couple. A good survey of research around explanations in machine learning can be found in [18]. One of the first efforts in this area [13] looked at explaining the decisions of classifiers in a model agnostic manner. However, a majority of subsequent work has been in explaining the decisions of deep learning models using various strategies such as saliency maps [8], influence functions [9], logical primitives [14, 15], and causal frameworks [10].

Interactive AI: Towards the goal of democratizing AI access, Google recently launched "AutoML Vision" [7], an AI product that enables everyone to build their own customized machine learning models without much expertise. In [22], researchers present a new system that automates the model selection step, even improving on human performance. Systems that can learn interactively from their end users are gaining importance. [20] is one of the early efforts in this area. While most progress has been fueled by advances in machine learning, the authors in [19] explore the notion of interactivity from the lens of the user. Recently,

model-specific interactivity is being introduced through efforts such as [21].

Tunable AI: This area is in its nascent stages. Services like Sagemaker and AutoVision claim to provide auto tuning facilities, but do not focus on the AI consumer.

TRANSPARENT AI FOR SALES "WIN" PREDICTIONS

Recommendation systems are an important class of AI applications in the enterprise. In the example below, we show how various aspects of transparency were essential in the adoption of an AI system for predicting sales "wins". This is an actual example of the process of selecting AI for an enterprise workflow; names have been anonymized.

The user who was trying out this predictive AI system was the global SVP of sales for a large enterprise company. Let's call her Allison. Allison used Business Intelligence dashboards custom built for her on a daily and weekly basis to *look at* various trends in sales data. The AI.AI platform made it easier for Allison to *ask* questions of the AI, and to receive answers as custom graphical representations with accompanying auto-generated text explanations. In this case, Allison's initial *ask* to the AI was:

How do I increase overall win % on sales contracts?

The AI answered:

Total contract price does, lower priced is better.

By means of an explanation, it provided graphical representations of contracts that were won vs. lost, with explanations.

A lower contract price as a winning sales strategy is not exactly music to a sales executive. Indeed, in this case, this particular recommendation immediately ran into a strongly held business belief of Allison's. A certain percentage of contracts were "churn" contracts, essentially contract renewals with low price but high "win" probability. The AI's response failed Allison's belief test, and her next ask was:

Hmm. Churn contracts (i.e., contract renewals) are affecting the result. Let's remove them.

And the AI's response:

Same result after removing churn contracts.

This led to Allison digging in:

Really? I wonder why. Show me the data that matches these conditions.

This was the beginning of an extensive series of edits that Allison performed using the platform to update the AI by asking it to look at a variety of subsets of the original data, asking it various questions along the way. The process ended once she arrived at an AI-driven insight: most contracts, despite not being coded as such in the dataset, had churn like characteristics. This was a huge insight at the level of a sales SVP, enabled because of her ability to *fluidly interact* with

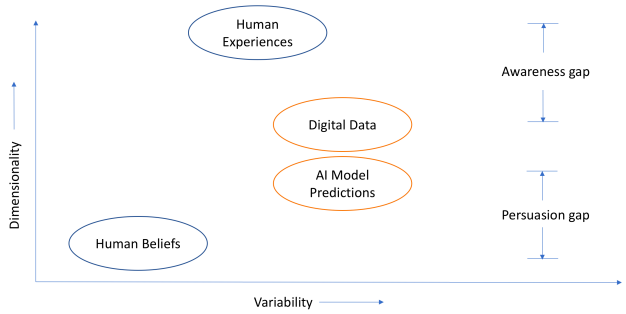

Figure 1: The Search for Common Cognitive Ground

the AI and pose questions and get immediate answers. Allison then asked:

What's the impact of total contract price on the remainder?

And received the answer:

Lower price is no longer better.

REFLECTIONS FOR AI TRANSPARENCY

Abstracting from the example described in the previous section (and other examples from our industrial research experience), we'd like to offer the following perspective, captured in Figure 1.

Human experiences tend to be highly dimensional; there are many aspects to the human experience. There is also variability to those experiences. Comparatively, human beliefs, which are borne out of human experiences, may be described as being lower in dimensionality as well as in variability. When we introduce digital actors, digital data, and digital decision making (AI), we end up at different points on the Dimensionality-Variability graph of Figure 1. Because digital data may not capture everything that is experienced, we may view digital datasets as having lower dimensionality than the data underlying human experiences. The predictions made by AI from digital datasets, may then be further lower in dimensionality, similar to the dimensionality difference between experiences and beliefs.

Two issues show up when a human being is presented with AI decisions. If they don't believe them because they do not align with their beliefs, they point to the lack of awareness of the dataset with respect to their experiences. Let's call this the *Awareness Gap*. The awareness gap is often used as a first line of defense to reject AI that offers no way to edit it, independent of its explainability features.

Similarly, if an AI's decision is not aligned with the user's beliefs, it is important that the AI be able to understand this gap and persuade the user by applying techniques from cognitive science. One issue we see in the explainability literature is too much of an implicit assumption that *rationality* is a winning persuasive argument whereas in reality this is far from the case. Closing the *Persuasion Gap* requires, in our experience, the ability of the AI to engage

mechanisms that human beings regularly use to update their belief systems, and recourse to rationality is only one such mechanism.

We suggest that research on the ability of the human actor to guide the AI to help it close the Awareness Gap, and on the ability of the AI actor to guide the human to help close the Persuasion Gap is essential for practical human-AI collaboration.

CONCLUSION

The justified excitement about modern AI has brought many people in non-technical roles in the enterprise into the sphere of AI interaction. Enterprises are re-architecting themselves to go from "Intelligences Apart" – human and machines intelligences being separate – to true human-AI collaboration. In many enterprises, incorporating AI into workflows goes through a pivotal stage of testing if it can work well with the existing human decision makers in that workflow. Human decision makers use alignment with their existing beliefs as a way of accepting AI into their team, much as they might for accepting a new human team member. For AI to pass this test, in addition to being explainable, it needs to be easily accessible and interactive. AI that is transparent in these ways can be edited usably by non-technical stakeholders when it fails their belief tests, and engenders trust in the process. In addition, we suggest that AI look to mechanisms from the cognitive sciences to both identify beliefs in their users and ways of updating those beliefs that leverage techniques in addition to rational explanation.

REFERENCES

1. D. Gunning. 2016. Explainable Artificial Intelligence (XAI). Retrieved December 20, 2017 from https://www.darpa.mil/program/explainable-artificial-intelligence
2. DataRobot. 2017. Automated Machine Learning. Retrieved December 20, 2017 from https://www.datarobot.com/
3. Z. Lipton. 2016. The Mythos of Model Interpretability, In *International Conference on Machine Learning Workshops*.
4. T.Millers et.al. 2017. Explainable AI: Beware of Inmates Running the Asylum, In *ArXiv Report*.
5. F.D.Velez, et.al. 2017. Towards a Rigorous Science of Interpretable Machine Learning, In *ArXiv Report*.
6. AWS RE:INVENT, 2017. Retrieved February 10, 2018 from https://techcrunch.com/2017/11/29/aws-releases-sagemaker-to-make-it-easier-to-build-and-deploy-machine-learning-models/
7. AutoML Vision, Google, 2018. Retrieved February 10, 2018 from https://www.cnbc.com/2018/01/17/google-launches-cloud-automl.html
8. R. Selvaraju et.al. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, In *International Conference on Computer Vision.*
9. P. Koh et.al. 2017. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning.*
10. D. Alvarez et.al. 2017. A Causal Framework for Explaining the Predictions of Black-box sequence to sequence models. In *ArXiv Report*
11. S. Tan et.al. 2017. Detecting Bias in Black-box Models Using Transparent Model Distillation. In *ArXiv Report.*
12. T. Bolukbai, et.al. 2016. Man is to Computer programmer as Woman is to Homemaker? Debiasing Word Embeddings, In *ArXiv Report*.
13. M.T.Ribeiro et.al. 2016."Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Conference on Knowledge Discovery and Data Mining.*
14. H. Lakkaraju, et. al. 2016. Interpretable Classifiers using Rules and Bayesian Analysis. In *The Annals of Applied Statistics* 9(3), 1350-1371.
15. T.F.Wu, et. al. 2016. Learning And-Or Models to Represent Text and Occlusion for Car Detection and Viewpoint Estimation, In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9), 1829-1843.
16. K. Muhammad et. al. 2016. Explanation-based Ranking in Opiniated Recommender Systems, In *Irish Conference on Artificial Intelligence and Cognitive Science.*
17. H. Lakkaraju, et. al. 2017. Learning Cost-Effective and Interpretable Treatment Regimes, In *International Conference on Artificial Intelligence and Statistics.*
18. O. Biran et. al. 2017. Explanation and Justification in Machine Learning, In *International Joint Conference on Artificial Intelligence.*
19. S. Amershi, et.al. 2017. Power to the People: The Role of Humans in Interactive Machine Learning, In *AI Magazine.*
20. C. Isbell et. al. 1999. The Parallel Problems Server: an Interactive Tool for Large Scale Machine Learning. In *Neural Information Processing Systems.*
21. Y. Hu et.al, 2011. Interactive Topic Modeling, In *Association of Computational Linguistics.*
22. T. Swearingen et.al. 2017. ATM: A distributed, collaborative, scalable system for automated machine learning. In *IEEE International Conference on Big Data.*