AI Is a Waste OF MONEY

Why Most AI Projects Fail and the Secrets of Succeeding with AI

Arijit Sengupta

First Edition: February 2019
Printed in the United States of America

To my two-year old daughter Audrey (Maya).
When you grow up and look at what AI has become,
I hope you will be proud of your dad.

Contents

Introduction

The vast majority of artificial intelligence projects fail. Currently, less than a third of all AI projects are implemented, but there is increasing evidence that the number is lower, closer to 5%. Why is that? And how can you make sure that your project is one of the few successful ones?

Some AI software vendors brag about the number of algorithms they support. Others focus on how accurate their AI models are.

That doesn't matter.

What matters is getting measurable business results from AI, making money or saving money at a scale that far outpaces your investment.

The media is no help here. Media coverage of AI is either fear mongering or hype. Both are dead wrong. AI is going to transform our way of life, even more than the Internet has, but very few people truly understand the power and limitations of AI.

Experts explain AI using mathematical models and academic terms instead of focusing on how we can use AI in the real world.

What we need now is a better understanding of how to use AI from a business point of view. I can show you how to succeed using AI because I learned the hard way what works and what doesn't.

I have dedicated almost two decades of my life to AI, starting with studying it at Stanford to selling my first AI company, BeyondCore, to Salesforce, and now starting my second AI company, Aible. I have invented AI-related technologies for which I have received seventeen patents, personally developed more than a thousand AI models for customers across many verticals, and helped deploy AI solutions at scale for some of the largest companies in the world. I regularly speak to business users about the realities of AI and have debates with my data scientists about the intricacies of algorithms.

I wrote this book to demystify AI for a larger audience and give you a practitioner's perspective. My hope is that by reading this book you will develop your own intuition and instincts about effective ways of working with AI.

At the Harvard Business School (HBS), I had to read over 800 cases over the two years of the MBA program. Each case laid out a business problem of some sort. At one point I asked one of my professors why HBS forced us to read so many cases. I still remember his response: "When these problems arise in the real world, you will have very little time to take the right decision. Here in the safety of your classroom, we are trying to help you think through the

possible scenarios and how you would react to them. This way, you will have developed your own rules of thumb that will help you take decisions faster if you actually encounter that scenario." His advice has turned out to be more accurate than I sometimes care to admit. I hope this book will help you think through the possible pitfalls of AI and how you would deal with them. Forewarned is forearmed.

WHY AI NEEDS HUMAN GUIDANCE

Both the power and the limitations of AI were demonstrated in a project in the early days of BeyondCore, the company I founded out of some research I did at HBS. We were testing an AI system designed to predict patient readmission rates. We got a lucky break when consultants from McKinsey & Company partnered with us. The McKinsey team pulled together health data on 30 million patients, and our AI program analyzed the data across a million variable combinations.

The AI immediately found interesting patterns. For example, it found that young women with diabetes had a significantly higher chance of being readmitted to a hospital when compared to older men with diabetes. This sounded strange because typically younger people have lower readmission rates than older people, and women are slightly better at taking their medicine than men. Thus, young women should have lower readmission rates than older men.

The McKinsey team was able to independently confirm the pattern in the data, but they could not find any reference to it in medical journals. In looking for an explanation, they

showed the finding to a few doctors who proposed a theory: that neglecting to take insulin was a quick way to lose weight. The team then looked at the evidence and found that indeed, young female patients—as well as a significant but lower proportion of young male patients—were deliberately neglecting to take their insulin. Without insulin, their bodies would not process the sugar they consumed and thus they would hopefully lose weight. Their desire to lose weight was so strong that they compromised their health and landed back in the hospital as a result.

Why hadn't medical researchers known about this pattern? They had simply not thought to ask this specific question. It made much more sense to focus scarce research dollars on older patients who they expected would have more readmission problems.

So why did the AI find this pattern? Because it simply asked *every possible question*—about a million in this case. This was simply one of those million questions. That is the power of AI.[1]

That project also illustrated the limitations of AI. The same AI strongly focused on a readmission pattern for another condition. This condition only occurs with women between the ages of 16 to 50 who were never readmitted for this particular condition for at least another 9 months.

[1] I presented these findings with McKinsey at StrataRx and you can see the video at https://youtu.be/UtRpC4er2CQ?t=1911.

What insight! The AI essentially figured out that only women get pregnant! While an AI may look at millions of patterns to detect which ones are important, it focuses on a small set of these patterns when it tries to predict. Because this was such a pervasive pattern, the AI fixated on it and wasted much of its predictive power on this pattern. But of course, the goal here was identifying avoidable costs and the AI had no way of knowing that pregnancy is a special medical condition that people may or may not want to avoid.

How is it that the AI identified a pattern that was brand new to top medical researchers while at the same time it focused on an obvious pattern that was unrelated to our objective?

Because AI has no context for the questions it is asking, it can be easily distracted by patterns that don't help you achieve your business goals. In this case, we had to exclude pregnancy-related data from the dataset and then retrain the AI.

Human intuition was a necessary component to make the AI effective. The interplay between human and AI is better described as Intelligence Augmentation as opposed to traditional Artificial Intelligence.

TREAT AI LIKE ANY OTHER INVESTMENT

When new technologies appear on the scene, there's a tendency to lose track of the fundamentals. AI is just like any other investment, but that's not how it's treated. Here are just a few aspects of this problem:

- AI projects are started without any expectation of business impact. And people are not even sure whether the AI they're creating is better than what they already have in place.

- No one tells you up front how much it will cost to train the AI. Data scientists regularly run 50, 60, 100, or even 1,000 models if they're using hyperparameter tuning. They have no idea how much that cloud computing bill will be. And the people who pay the bill don't track it down to see who has blown the budget.

I regularly see companies spend millions on AI projects with no projected business impact and no commitments about their costs. That is no way to do business.

Suppose I told you that I have magic water in a bottle. I can't guarantee that my magic water will cure cancer, I have no research on its success curing cancer though it helped my friend cure tuberculosis, and I can't tell you how much it costs. Nevertheless, you need to commit to spending lots of money on this magic water *right now*. In no other context would you accept that deal. But everyone is so excited about AI that projects are run like this all the time.

You need a way to benchmark your AI projects and get a sense of expected return on investment (ROI). All of the same principles of business investment still apply.

FOCUS ON PRACTICAL AI THAT YOU CAN USE TODAY

There are many types of AI technologies, but in this book, I have focused on use cases that a typical business will deploy. Very few of us will design self-driving cars, but many of us will work with AI to predict trends and inform decisions.

I define AI through the lens of practicality:

> *AI is a way for software to automatically learn and codify useful patterns by looking at examples from the past. These patterns can then be used to explain why things happened in the past, what is likely to happen in the future and what actions we can take to affect future outcomes.*

Note that I don't use the word 'thinking' in my definition. If AI clearly learns, does it matter whether it thinks or not? Now, some experts will argue that my definition applies solely to machine learning and not broader AI. But my definition most closely explains AI as experienced by business users today.

Continuing with the theme of practicality, I find discussing real examples and use cases is the best way to bring you up to speed quickly. I have many exciting stories to share, with practical takeaways. But it would be nearly impossible to get approvals from hundreds of customers to tell their stories. As a result, I have fictionalized, anonymized, and blended real customer stories into the adventures of a newly minted

data scientist named Vera and her Harvard Business School-educated mentor Jit.

In this book, I have tried to cover most things you need to know to get a better grasp of this technology. Here is my most important message:

> *AI needs you: Your expertise, your common sense, your business acumen.*

If you want to be ready for the next evolution in our society, you need to understand the contours of AI's power and weakness. That is why I wrote this book.

Let's get started.

Section I

What Can AI Do for Us?

The power of AI comes from the fact that it can provide us with actionable information at the moment of making a decision. Before AI, decision support or BI systems would present us with complex visualizations about what happened in the past and the onus was on us to take the right decision based on that complex information. Now AI can tell us exactly what is likely to happen in a very specific context. In sales, it can tell us how likely it is that we will successfully sell to a specific customer. In marketing, it can tell us the probability that a customer will open a marketing email. In support, it can tell us the likely customer satisfaction with a call even before it has been completed. In operations, it can tell us the likelihood that a part will fail. In fact, several academic books define AI as a tool for predictions, but in reality, it is so much more.

In the following four chapters, I outline how AI is used for predictions and the best practices for introducing AI into business processes. But you will also learn about how the patterns an AI learns are valuable in themselves. When an AI is trained on your data, it learns a lot about how your employees work. For example, it may learn the kinds of deals your salespeople pursue aggressively and the kinds of deals they avoid. When executives review such patterns they often respond that there is fundamentally nothing new here. They already knew what the AI has learned. This completely misses the point. Of course, the expert executives who know their business extremely well might be aware of such patterns. But here the AI independently learns such patterns from the data itself and it can now disseminate these patterns at scale to all of the employees.

Imagine a world where your newest employee has almost the same context-specific knowledge about your business that your most experienced employees have. AI can be a wonderful way to build up institutional knowledge about your organization's best practices. As part of this process, you will also invariably learn something new about what your organization is doing. These nuggets of previously unknown insight can often pay for the entire AI project.

You will also start to appreciate the importance of the human element in creating an AI. Instead of thinking of AI as a means of automating humans, you will see how feedback from experts is a crucial element of creating a successful AI. You will also learn to detect when an AI is good enough for the job. You are better off adopting a 'good enough' AI that creates clear ROI today rather than spending time trying to create the hypothetical perfect AI.

But AI can offer so much more than predictions and explanations. In its most useful form, AI is used for generating context-specific recommendations. If old-school BI is a printed roadmap, then AI-powered recommendations are like driving with a GPS with live traffic updates and turn-by-turn directions. In this analogy, predictions are like knowing whether or not you will be late for your next meeting. It is certainly useful, but not as useful as knowing how to avoid the traffic so you can actually make it to the meeting on time. While we touch upon the power of recommendations in this section, we will cover it in more detail in Section II.

Chapter 1

Welcome to the Real World

In which Vera finds out that data science courses did not prepare her for the world of business.

It is my first day at the customer site, and I'm ready to quit. Here I am with a degree in statistics and machine learning from Carnegie-Mellon, and my first gig with Foundation Consulting? Building some sales dashboards. BI 101.

ManuCo, a large industrial parts manufacturer, hired us to help them improve sales. My project lead Morgan and I had a meeting with their CFO, Todd Clemens.

Todd launched right in. "I need my salespeople to know absolutely everything about the customer before they go into a meeting. And I want my sales leaders to have a coaching dashboard where they can see how each salesperson is performing compared to their peers."

Please tell me it's not my job to make sales dashboards. I glanced sideways at Morgan, who is listening attentively to Todd and taking notes.

There is so much more we could do here! I had just aced a great course on sentiment analysis. This kind of AI is used by so many leading companies like Google, Facebook, and Apple. I was sure it could help ManuCo as well. I explained that we could tell him how customers think about his products based on a cool sentiment analysis of what his customers said about ManuCo products on Twitter and Facebook. But from the look on his face, I got the distinct impression he had never used Twitter.

Morgan shot me a look and said, "Vera will have the dashboards ready by the end of the week, Todd."

The very next day, I went to Morgan with my concerns. She shrugged. "Talk to your mentor. Didn't you meet with him last week like I told you to?" "No," I admitted.

Jit, my 40-something mentor at Foundation Consulting, has an MBA from Harvard. Does he know anything about AI? Not if this place is any indication.

To get Morgan off my back, I emailed Jit, and he offered to take me to breakfast to talk about the project. Maybe he can get me a better assignment, and give this dashboard work to someone who really likes it.

After a brief getting-to-know-you chat while we were waiting for our breakfast, I complained to Jit about how Todd was completely clueless about AI and didn't give me any time to explain how we could apply sentiment analysis for him.

After listening to me for a while, Jit asked me what ManuCo actually sold. I had to do a quick Google search on my phone because to be honest I hadn't considered the specifics of their product line before. Turns out they sell things like ball bearings and metal parts. What a boring business!

Jit then asked me, "Do you think ManuCo's customers ever tweet about ball bearings?" I had to admit that I was so excited about the methodology that I had not considered whether it really applied to the customer's problem.

We agreed that we had to give Todd a better vision for AI but before we could focus on that, Jit wanted to ask me some questions. "What exactly is the customer trying to achieve? What is their goal? Why did they start this project?"

The truth was I didn't know the answers, and Morgan had gone on vacation, leaving me to work directly with Todd and his team. I said I'd ask the customer.

Since we knew the problem was related to sales, Jit gave me some more questions to ask and ran me through a few examples of how things go wrong. He urged me to listen carefully to the customer's responses.

When I asked Todd why he hired us, he explained ManuCo's problem. "Our profits have been dropping over the last few years even though we are closing more deals than ever. If we can't fix the problem, we are all out of a job. That is why I am trying to enforce better sales discipline."

Seemed to me like he jumped a few steps there and made some assumptions. Why does he believe that better sales discipline will solve the problem of profits dropping? If salespeople are closing more deals than ever, many other factors could be affecting profits. One thing I know from training AI models is that assumptions have a nasty way of biting you in the ass.

I was going to have to drill to get the answers from Todd, and despite talking to Jit, I didn't understand the business well enough to ask the right questions. I thanked Todd and told him I'd get back to him.

Jit and Todd were both open the following afternoon, so I got a meeting for us. I was mortified I'd already had to bring in Jit, but I needed help, especially with Morgan out of town.

Jit asked Todd to explain what drives profits at ManuCo.

"It depends," Todd said. "Sometimes we sell directly, sometimes through dealers, and some of our biggest deals are driven by partners who order a custom component to incorporate in their own product. Actual profits for those custom deals depend on factors such as pricing and production volume."

"If you really want to improve your profits," said Jit, "you'll need to answer some questions:

- What impacted profits in the past?

- Why and how did these factors impact profits?
- What are the expected profits for the future?
- How can we improve profits?"

"That's why I hired *you*. Those are the types of questions I need to be able to answer. I need new sales dashboards," said Todd, glancing at me.

Jit looked thoughtful. "I know what you mean about dashboards. They can be great for senior executives who have a strong feel for the business, but I think we can do better."

"What do you mean, better?" said Todd.

"You mentioned that you want dashboards for your salespeople. Let's think about how we can get them the most effective information. Let me ask you this: are you worried about the past or the future?" said Jit.

"The future, obviously," said Todd. "If we don't increase our profits, we're out of business."

"Dashboards are for people who want to quantify the past and see what happened. They're backward facing, like driving a car using the rear-view mirror," said Jit. "Your strategic initiative for this quarter isn't to have a prettier dashboard, is it?"

Todd looked uncomfortable. “Of course not.”

“The other problem is that dashboards often hide the truth,” Jit continued. “They look like you can extrapolate from them, but they aren’t at the right level of granularity to support that.”

“What do you mean?” asked Todd.

Jit started drawing on Todd’s whiteboard. “Here’s a simple example. Imagine you have 100 sales, and every sale has a profit of $100. The total profit is $10,000.”

“Now let’s look at another scenario. One sale has a profit of $9,010 while 99 sales each have a profit of $10, so again, the total profit is $10,000.”

Number of Sales	Profit Per Sale	Total Profit	Average Profit
100	$100	$10,000	$100
100	1 sale at $9,010 99 sales at $10	$10,000	$100

“On most dashboards, you would either graph average or total profits, so these two scenarios look identical. That’s one of the fatal flaws of traditional dashboards,” said Jit.

Jit just calmly kept going with this anti-dashboard explanation.

I was working hard to keep a smirk off my face. I really wondered what Morgan would say if she were here. Still, Todd was listening, even though he clearly had questions.

Jit continued, "The other fatal flaw of dashboards is even more insidious. When salespeople look at graphs, they make assumptions. If a graph shows that average profits in Texas are $100 higher than in California, they naturally think profits in general are higher in Texas. But what if the results in Texas were driven by one exceptional deal?"

Todd objected, "But I remember there is a way to check for patterns like that. It has been a while though since my last Statistics class."

"Of course, it's possible to see this with calculations like variance or standard deviation," said Jit. "But people who design dashboards rarely conduct such tests, and business users looking at the dashboards don't have what they need to calculate such metrics. This leaves the door wide open for business users to extrapolate what they *think* the data is telling them."

"When salespeople are at the moment of taking action, you don't want them to look at 15 graphs to figure out what they should do," said Jit. "It's better to give them something more specific. They need guidance, informed by the specific context and ManuCo's priorities, that says, 'Here is what you should do. Here is the next best option. And here's the third best option. Choose one."

"Essentially, at the time of taking a decision, you want to give them context-specific advice," Jit continued. "Graphs are not context-specific. By definition they are generalized. You have a graph of revenue by product, revenue by geography, and revenue by salesperson."

"But that works for me," countered Todd.

"But does it work for your salespeople? If you want to empower salespeople, you need to give them predictions and recommendations at the moment of making a decision, because they don't have time for anything else. You don't want them stepping back and thinking about graphs. You want them in the zone closing deals."

"It sounds like you're talking about AI," said Todd.

"I am. Predicting the future is where AI shines. But what's a little harder to understand is how AI does this," said Jit. "AI picks up on subtle statistical patterns that humans miss. For example, if a salesperson does badly in general, you would probably know about it. If you are doing badly with a vertical, you would probably know about it. But what if one salesperson was doing worse than others in a specific vertical? Or let's say the salesperson is doing very well, but given their territory and the tenure of their customers, they should be doing 50% better? Would you catch patterns like that?"

"I see your point, but it still makes me nervous. Where are these recommendations coming from? How can I validate those next actions?" said Todd.

"Validation is crucial," said Jit, "but iteration is even more important. The key to success is understanding that AI isn't going to make perfect predictions. Rather, it will quickly present a set of options that would be better on average than any that you would get, except from your best people. Your sales staff does the validation by using their knowledge of the market and the customer to pick the option that makes sense or reject them all. They are the augmentation of the AI, the crucial link to quality. The AI can in fact learn from which predictions the team acted upon and which they ignored, to get even better over time."

Todd seemed both energized and deflated at the same time. He said, "OK, why don't you guys take a few days to show me what is possible and then I can decide."

I never did create those dashboards. Instead, I worked up a model to predict expected profit and probability of success for each sales opportunity that the ManuCo sales team looked at. ManuCo started using this information to prioritize leads and predict future profits. The model was working! With a little help, I knocked my first project out of the park.

JIT'S TAKE

As Vera's mentor, my mission is to speed her transition from the "assume we have a can-opener" world of academia where datasets are pristine and questions are well formed to the messy real world. In most projects, the data is a mess at first, and there is little consensus on the most important

business problem until someone asks clear questions that provoke a frank discussion.

Nonetheless, this was a good first project for Vera, and for the customer too. Customers all too often jump directly to the solution before fully discussing the business problem they want to solve.

In my view, AI is the new BI. BI dashboards provide general guidance and a sense of history. AI offers specific next actions at the point of decision. That's why BI dashboards alone are rarely the right answer nowadays. They made sense when it was impossible to provide actionable insight to the end user at the moment of making a decision.

Think about it: Would you rather know three things you can do *right now* to increase the probability of winning a deal, or would you rather see a graph of win rates across thousands of deals, only a small percentage of which are relevant to the decision you are trying to take right now?

Myth

Dashboards tell me all I need to know.

Reality

Dashboards are not actionable for the majority of users.
AI is the new BI.

Chapter 2

AI for the AI-resistant

In which Vera learns that even a business completely resistant to AI predictions can still profit from AI.

Jit and I are in the midst of a pretty strange project at a leading retail bank, BigBank. Liz Owens, their analytics manager, started the meeting by telling us why she would never use AI for making predictions

"You are here because our CEO is interested in AI," said Liz. "But we are in a highly regulated industry and we have to get approval for our predictive models. There is no way we are going to use AI for predictions."

"Have you ever tried using AI for predictions?" I asked.

Liz looked impatient. "Oh yes. And the last time we tried, it backfired. We did our best to avoid using variables like gender and race but the AI kept using them even when we took those variables out of the data. I could not understand what it was doing, and even more important, I couldn't explain it to the compliance team. We just can't use AI."

Liz raised a concern I've heard a number of times, even

though I've only been here a few months. AI seems like a black box; even experts have trouble figuring out why it's doing what it's doing sometimes. It's the exact opposite of transparency (or can be).

Because AI looks at so many factors, it can often figure out proxies for variables. For example, because more women are teachers or because women typically live longer than men, an AI can pick up hints of a person's gender based on their profession or even their age.

Essentially, if gender is a good predictor, and that variable is not available, the AI will ferret out other things that are proxies for gender and use those instead. There are ways to address this characteristic of AI. It is somewhat complex, but it can be done.

But this is only one symptom of a bigger problem with black-box AI: you don't know what it's learning. It can identify a pattern that is completely wrong without anyone realizing it.

Reversing Cause and Effect

There's a famous cautionary tale about an AI that was being used to triage patients to determine which patients the doctor should see sooner. The doctors started noticing something strange about the predictions because the AI was prioritizing certain risky patients lower than it should have. It took them some time to figure out exactly what was going on.

If a patient had diabetes, the AI prioritized the patients lower than the doctors would have. It turned out that the doctors were especially careful about treating riskier patients like those with diabetes. As such, their treatment took a little longer and the immediate health outcomes were a bit better than usual.

To the AI, it looked like these patients required more time and had lower risk (because the immediate health outcomes were slightly better), and so it started prioritizing these patients lower.

Of course, any human would have realized that their outcomes were not better because of diabetes; they were better because the doctors prioritized these patients higher because of the additional risk.

The AI confused cause and effect, and thus learned the wrong lesson.

The standard way to address these types of concerns involves having the AI explain the basis of its predictions to the user. But, in this case, Liz did not seem open to that discussion.

Jit was sympathetic. "We completely understand. Why don't we use AI to look for patterns in your data you might not have seen before? We won't predict anything. Think of it as a double check or audit of your data."

I added, "We'll deliver some slides similar to what your analysts create today. The difference is that the AI will have

evaluated millions of patterns instead of the dozens that human analysts evaluate manually. This way your CEO gets an AI project while you don't have to worry about regulatory oversight problems."

Liz looked skeptical, but gave us the go-ahead.

We started by analyzing customer satisfaction and attrition data. For each customer, we included everything we knew about them, their customer service interactions, and whether they had closed their accounts.

The AI quickly identified certain patterns that helped us convince the bank that it was learning the right kinds of lessons. For example, it noticed that if the customer started making large transfers of funds to a different bank, that was a good predictor that they were planning to leave. That made sense to the bank, and they had pretty much always known this pattern.

The AI also found some patterns that the bank did not know about. For example, it turned out that if the customer was a small business with at least a $1,000 line of credit, then they would almost never churn. This would have been a great factor in a predictive model, but we were not allowed to do predictions. So we wrote this up and passed it to Liz (along with the analytical proof).

Surprisingly, even Liz was excited about our findings. "If this is true," she said, "we are better off giving every small business customer a small line of credit instead of spending

all the money we do on customer retention programs."

Essentially, instead of using AI to predict the probability of churn for each individual customer, she was willing to fundamentally change her business to systematically reduce churn. That kind of action can have a huge impact on the business. No predictive model required.

The analysis also revealed that individual customers who were automatically paying at least eight bills a month from their account were very unlikely to churn. The bank had previously known about this pattern, but they did not realize that the effect really kicked in only after the customer had started auto-paying eight bills. The bank had previously pushed customers to sign up for their autopay program, but it was not having the hoped-for impact on customer retention. Now, they shifted the focus to increasing the number of payments made automatically.

For a project that started on a pretty low note, we were delivering quite a lot of value. The best outcome, though, was a change in the bank's attitude about using AI for predictions.

They informed us that several of the patterns we found were so useful that they were incorporating them in their manually crafted prediction models and putting them through the regulatory approval process. Sure, it wasn't a model that I created, but at least it was *influenced* by an AI model I created. As long as we impacted the business, the means matter less than the outcome.

JIT'S TAKE

AI is not monolithic and it certainly does not have to be a black box. But the marketing departments of AI vendors often position it as magic, causing a backlash when customers react to what they think AI is instead of what AI truly is.

The easiest way to address this is to step back, create tangible value using specific insights, and even the strongest detractors turn around because they realize that what you are showing them is very different from what they imagined. In fact, I often prefer to use the term Intelligence Augmentation (IA) instead of Artificial Intelligence precisely because of the mythos surrounding AI. Liz was not open to an AI making predictions. However, she was willing to examine the patterns the AI had identified and see the merit in them. In this way, the AI helped Liz, augmenting her intelligence with additional information that she could digest and act on.

I would suggest that nearly all businesses can benefit from this type of use of AI.

Myth

AI is all about predictions. If I'm not ready for predictions, I can't use AI.

Reality

AI is great at identifying patterns in your data that will help you better understand your business processes. Some of those patterns will drive value.

Chapter 3

People Matter

In which Vera learns that cutting-edge algorithms are not the key to success.

Jit and I have been called in to help a really big company this time: BigRetailer.

Aretha Ramirez, their CFO, wants help reducing shrinkage, which is taking a multi-billion dollar bite out of their profits each year. Shrinkage seems to be polite industry speak for theft or damage of the items they sell. We're meeting with Aretha and her VP of IT, Clint.

Aretha kicked off the discussion. "We've added inventory tagging technology to reduce shrinkage in certain categories like electronics, but we have such a large assortment of products in our stores that it's hard to see where we're being hit hardest and come up with effective solutions to reduce those numbers," explained Aretha.

"Even more complicated is the fact that shrinkage could be occurring along the supply chain, in the warehouse, on the trucks—anywhere," she said.

"I've heard that data is the new oil, and we have lots of data," she said, nodding at Clint. "From what I understand, AI on all that data should really help us."

Clint looked a little agitated at this point. "Data may be the new oil, but if so, I need some time to get it out of the ground and refine it before you two come in and start training an AI on it."

"We don't have the right systems in place to bring the data together. My view is that we should spend a couple of years building out the right infrastructure. You need more data, clean data, and comprehensive data if you really want to see what's happening across the business and contributing to the shrinkage problem. AI needs data; I think everybody agrees on that," said Clint.

"Actually," I said, "did you know that you can actually do more with less today using techniques like AutoML that focus on using the best algorithms? You don't need as much data and you don't need data that's in perfect shape. Yes, more data and cleaner data would be great, but we can get a good start with what you have. Better algorithms can actually be trained on smaller amounts of data than we had to use in the old days when we were just using statistical methods."

Clint was unconvinced by Vera's argument. "I'm not sure you appreciate how many data sources I'm talking about, and the fact that they are not unified or consistent in format. We have point-of-sale systems, inventory systems, logistics,

purchasing, marketing, and more. Then we have data being exchanged with thousands of vendors and partners via APIs to bring a million plus items into more than 1,000 retail outlets across North America."

"But Clint, we have been able to produce results for our clients using much less data than you'd expect," I said. "The latest algorithms can help you get started."

"With all due respect," Jit said, "I disagree with both of you."

I looked at Jit, surprised that he was disagreeing with me.

"I believe the human element beats better algorithms as well as better data," said Jit."

Clint looked at him. "Are you saying better data scientists deploy better models and collect better data, so investing in top data scientists is more important?"

Jit responded, "Not exactly. Let me tell you a success story that involved no data scientists but where the human was more important than anything else."

"This involves one of my favorite clients, Jonathan, who was also a VP of IT like you. Jonathan was tired of waiting around for the data scientists to create a model that predicted which sales were most likely to close. The project was key to the profitability of the company and it had stalled completely while the data scientists cleaned the data."

"While he was only supposed to provide IT support to the project, because he knew the data well enough, he decided to experiment with it. He started with the same data that the data scientists were using and created a simple predictive model with it."

"Jonathan then started predicting the probability of winning deals for live data coming into his system, but did not share the predictions with anyone."

"The model did fairly well. But he noticed certain types of transactions where the model predicted a low probability of winning even though expert salespeople seemed to be successful at closing those deals while newer salespeople were not."

"He talked to the expert salespeople and learned that they pursued deals where the customer had attended the annual user conference, which they viewed as a key indicator of engagement and commitment."

"User conference attendance was not in the original dataset, so Jonathan added that variable. The predictive accuracy went up significantly."

"He kept looking for similar patterns, talking with business experts and iterating his models based on what he learned from them."

"At the end of two months, he found that he had run dozens of models while the data science team was still debating data

quality and which algorithm was better."

"The company ended up operationalizing the model created by a curious and motivated user with no data science background. Jonathan was not a statistician, but he was an expert at looking for patterns and talking with people to learn more. He was systematically extracting the domain knowledge that was stuck inside the heads of the most knowledgeable business users and operationalizing it in the form of AI."

"That human curiosity and stick-to-it-iveness trumps data and algorithms any day. Motivated business users like him, not dry algorithms or data, are the reason why I do what I do and why we as an industry will succeed," Jit said.

Aretha looked encouraged. She had risen through the ranks over the years and took pride in her deep understanding of the business even though she was less confident in her understanding of things like AI algorithms.

Jit said, "AI's real role is to find patterns that we use our domain knowledge to explore further. It helps us. It augments our intelligence; it doesn't replace us."

Clint looked thoughtful. "I think I know who you should work with," he said. "George, our comptroller, is one of the most curious people I know. He will be great at helping you track down shrinkage and figuring out who to talk to," said Clint.

"Come to me when you are ready to look at some data and we'll figure out where to start," he said. "Meanwhile, Aretha, I'm still going to work on bringing all of our data together. We need that."

On our way to lunch, I said, "Jit—I didn't want to debate you in front of the client but I think you are underestimating AI, and it's only getting smarter every day. Didn't you see Deep Blue beat Kasparov and Watson win Jeopardy? Machines beat humans all the time. You are living in a world of humanist fantasy."

Jit responded, "I am not sure that we learned the right lessons from the Deep Blue-Kasparov game. Some commentators believe Deep Blue won because of a buggy move that distracted Kasparov, who did not realize the move was due to a bug."[2]

Naturally, I walked right into a topic where Jit knew the story better than I did.

"More importantly, Vera" said Jit, "I am not claiming that the human is superior to the machine. I am just claiming that the human and machine working together trumps human *or* machine."

"One example is the game of Advanced Chess that Kasparov himself helped create where the human player can use a

[2] "Did a Computer Bug Help Deep Blue Beat Kasparov?" https://www.wired.com/2012/09/deep-blue-computer-bug/.

computer to plan out chess moves. Typically the human-computer team is able to beat both humans and computers playing alone."

"Interestingly though, the best Advanced Chess teams are not the ones with the most powerful computers or the highest rated grandmasters. The teams that do best have figured out how to leverage the strengths of the human and the machine in the most synergistic ways,"[3] Jit concluded.

Hmm. I think I see what he is getting at. This is again about the value of human guidance when the AI is learning the

[3] "In 2005, an advanced chess tournament took place that allowed any combination of humans and computers. Steven Cramton and Zackary Stephen, who only held amateur status in Elo (named after physics professor Arpad Elo) chess rankings, took their regular desktop computers and squeezed them for their purposes. They won that tournament against chess masters with superior chess ratings and even superior hardware and software. Both players had leveraged their expertise to align computing power to win chess games. They created a superior team comprising humans and machines. In essence, a new form of chess intelligence had emerged. Kasparov concluded, 'Human strategic guidance combined with the tactical acuity of a computer was overwhelming.'"
From computer to centaur—Cognitive tools turn the rules upside down http://www.kmworld.com/Articles/News/News-Analysis/From-computer-to-centaur---Cognitive-tools-turn-the-rules-upside-down-101525.aspx.
Also see:
https://www.huffingtonpost.com/mike-cassidy/centaur-chess-shows-power_b_6383606.html; http://www.bbc.com/future/story/20151201-the-cyborg-chess-players-that-cant-be-beaten.

wrong things from the data. Reminds me of a cartoon I saw about how the AI apocalypse was an easy win because the robots used bows and arrows—because they learned that across all human wars, bows and arrows were the weapons of choice. The AI is powerful, but when the human helps it learn from the right data and protects it from drawing the wrong conclusions from the data, then the AI is even more effective. Why didn't he just say that?

JIT'S TAKE

The story of Advanced Chess is a useful illustration of a far more important point. Over time, it is entirely possible that a machine would also be able to consistently beat the combination of a human and a machine. This is because chess is a game with simple rules and clear objectives where human fatigue and error plays an important role. The world of business is not so clear-cut. If a business could be perfectly expressed in the form of simple rules, we might really not need humans. If business did not change over time, we might not need humans. But in a world that is complex and ever changing, I will always bet on the power of human domain knowledge in conjunction with the machine's power of math at scale.

Myth

1: To win at AI, you need lots of data.

2: To win at AI, you need the best algorithms.

Reality

To win at AI, you need curious people who know the business well to iterate and help make the AI better and better.

Chapter 4

Good Enough for the Job

In which Vera realizes that there is a point of diminishing returns, and she can't always get an A on an AI.

Usually, I am the one obsessed with the accuracy of a model, but this time it's the client who wants a level of accuracy that simply may not be possible for their business. Susan Nathan is ChemCo's new Head of Sales Analytics and she wants our model to be even more accurate about predicting the likelihood that they will close a specific deal.

When we first started working with ChemCo, their data quality was pretty bad. For example, most of the time, the competitor field was left blank in their customer relationship management (CRM) system, and reports run on the system showed that the most common competitor was No Competitor. The running joke inside the sales group was, "We have no competitors, so we lose to no one!"

Why would salespeople take the time to document such information if they don't see any direct benefit from it? This is just simple human nature. As one salesperson told me,

"Recordkeeping doesn't help us make a sale; it just gives management ways to beat up on us if the deal falls through." As salespeople started seeing the results from our AI though, their tune changed a bit.

Although the competitor field is often blank, ChemCo actually has a strong competitor called InjectCo. If InjectCo is in the mix, generally ChemCo loses the deal. Things had gotten bad enough that ChemCo salespeople simply stopped working on a deal if they knew InjectCo was involved.

Our AI independently learned that fact from the data and accordingly predicted lower probabilities of success and lower profitability for deals where InjectCo was in the mix. But, as usual, the AI noticed patterns no one else saw. Contrary to the overall losing pattern, in the oil and gas industry, ChemCo actually tended to win against InjectCo.

The AI thus rated InjectCo deals in oil and gas as high priority. Because the salespeople had strong financial incentives for closing highly rated deals, they ended up pushing hard on these deals—and they won against InjectCo! The AI actually helped them figure out a pattern that they could use to make money, which in turn increased their acceptance of AI.

But if sales wanted the AI to find more moneymaking patterns for them, they needed to fill out the competitor field. All of a sudden, we hit 80% completion rates for the competitor field for new deals. Several salespeople even went back and added in the competitor for old deals they had already won or lost.

Data entry went from a chore that might lead to criticism on lost deals, to a way to make money. As they put in better data, the model improved, eventually predicting whether deals would close with about 75% accuracy. Not bad given the complexity of ChemCo's business and their fair-to-middling data quality.

But Susan still wasn't happy. "When I was an analyst, I regularly hit 85% accuracy using Excel. AI should be able to do better, right?" she said.

I couldn't figure out how to respond to her claim that she could do better in Excel. "Let me get back to you tomorrow," I said, on my way out the door.

I asked Jit for advice at our breakfast meeting the next day. As usual, he answered my question with another question. "Have you asked her how she actually uses the prediction?" he said.

"First tell me how to respond to the Excel accuracy question," I demanded. "My model can do better than a spreadsheet, right?"

"Of course your model can do better, but should it? Will your additional work, which she is paying for, pay off for her? That's why I need to know more about how she uses the results," said Jit.

"She uses the prediction to determine which deals are going to close," I said, frustrated (more with Susan than Jit, if I'm being honest).

"Yes, but exactly which business process uses the prediction, and how does it use it?" said Jit. That I did not know.

When I asked Susan, she explained that deals were handled in one of four ways:

- **Highly likely sale and very profitable:** Hand it to sales and to make sure they pursue it, give salespeople extra incentives to close such deals.

- **Potential sale so have sales evaluate it:** If a deal has a greater than 60% probability of closing, it is handed directly to sales.

- **Maybe someday but just nurture for now:** If it has a 30–60 % probability of closing, it is given to marketing for follow up.

- **Unlikely to close:** Deal is recorded in the CRM system, but no specific action is taken.

The truth is that measuring model 'accuracy' is complicated. There are many different ways of measuring how good a model is,[4] but the amount you invest in increasing accuracy must pay off to make it worthwhile.

Let's say that an AI tells you that a deal has a 61% chance of

[4] For more details, read analyst Dean Abbot's blog (https://www.predictiveanalyticsworld.com/patimes/defining-measures-of-success-for-predictive-models-0608152/5519/), excerpted from his book, Applied Predictive Analytics (Wiley, 2014).

closing. If it's wrong, you'll be disappointed. If an AI tells you a deal has an 89% chance of closing, and it's wrong, you might have higher expectations, and be upset. But if your business process treats both of these deals as identical, precise accuracy metrics don't matter.

And at ChemCo, a deal with 61% and 89% probability of closing were treated exactly the same way, so the impact on the business of getting either of these predictions wrong would be identical.

We *could* push harder to make the model better, but that would likely have no impact on the business outcomes for ChemCo given their operational processes.

Now we face a question: Should we do more work for this client even if it won't really benefit them? This was something Jit had to handle, so I set up an appointment with Susan and Jit the next day.

"Susan, your question boils down to ROI," explained Jit. "We *could* continue to work on increasing the accuracy of the prediction. But, given the way your business works, we are not sure it would help you make more money."

Susan looked skeptical.

"I have a good way to explain it," I said. "Imagine you're in a class where the professor has specified that if you score above 75% on your final paper you would get an A. Now imagine you create an AI to predict your grade. If predicting

the letter grade is all that matters to you, do you really care about precisely predicting your raw score for the final paper as long as it correctly predicts your letter grade?"

"Your model is doing all you need it to do. There's no sense paying for additional accuracy that won't pay off," I said.

"That's right," Jit said. "This solution is good enough for the job you've given it. But if you want us to continue working on it and keep taking your money, we will be happy to oblige."

When Jit put it that way, Susan decided to declare victory on the project. I guess she had other places to spend her money.

JIT'S TAKE

I was proud that Vera managed to focus on the business objectives rather than giving into her competitive drive for predictive accuracy. The danger with a number like accuracy is that because it is simple to communicate, it is easy for people to focus on it to the exclusion of everything else. This is like a teacher who focuses on test scores instead of whether the student is actually learning the material and getting the necessary skills.

It is important to remember that the business outcome is the goal, and the model accuracy (or test score) is merely a way to measure a part of your efforts to achieve that goal.

Myth

The more accurate your AI model is, the better it is.

Reality

Don’t invest in more accuracy than you need to reach your business goal.

Section II

How to Select Models and Deliver ROI

So, you have been charged with running an AI project. How do you select the right AI and ensure you deliver a successful project as determined by ROI? There are some fundamental landmines you need to understand.

Accurate models can help you lose a lot of money. I was working with a mortgage insurance company that was very proud of the fact that they rarely ever had to pay a claim except in the case of death or divorce. They had deployed an AI that was very conservative and had seen claims on new policies drop drastically. Is that necessarily a good thing? Well, in this case the company had started passing on insuring so many good mortgages that the cost of missed opportunities had grown far higher than the expected loss from claims. In AI, accuracy can be defined in many different ways, and just because a model is accurate doesn't mean it is profitable.

Don't undervalue manual processes. Data scientists often calculate how much more accurate an AI is than random chance to demonstrate that the AI has value. The problem is that manual business processes are not run based on coin tosses. Your organization over time has built certain manual business processes that work at a certain level of efficiency. For example, salespeople have always qualified sales opportunities based on their likelihood of being successful. An AI offers a different and potentially more effective way of estimating the likelihood of success. But it is insulting to act like the expert salesperson was just tossing a coin to decide which opportunities they were going to focus on. If

you undervalue your existing manual processes, you will overestimate the expected ROI of using AI.

Realize that AI is fallible and will need to be updated over time. Don't treat it as magical or you will be setting yourself up for failure. You will succeed only if you frame your AI project in terms of continuously learning and iterating. You will also need more than one model in many cases. This is not a problem. Think of it like a specialist such as a cardiologist as opposed to a general practitioner. You need both.

Traditional pilots are a terrible way to evaluate AI. The underlying principle of a pilot is that if the software worked for a limited time for a limited scope, it will work over time on a broader scope. Well, AI is very context specific. For example, you can't train an AI on one country's data and then use it in another. As such the traditional pilot approach doesn't work well for AI.

Understand who you are trying to assist with the AI. For the majority of users, especially inexperienced ones, you might be more successful by providing recommendations that are easy to follow. On the other hand, if you are using AI with more experienced employees, you might be better off just offering a prediction that gives the users flexibility in how they use the predictions.

Chapter 5

Accuracy Isn't Everything

In which Vera learns that a model can be both accurate and useless. And that she needs to ask a lot more questions.

Since our last engagement, ManuCo has gotten more into modeling on their own, and now they're running into problems. I've been brought in to help their data science team. This should be a lot of fun because I get to work with experts as opposed to business users for once.

They created a model to predict the probability of customer churn. As I reviewed it, I was impressed; their model has really high accuracy.

But business users are complaining that it doesn't help them actually *reduce* churn. They say something strange is going on.

I have seen this movie before. The problem may not be with the AI at all, but with people who are closed minded about AI.

Still, if there's a problem with the model, I should be able to figure it out pretty quickly.

It turns out I spoke too soon.

I double-checked the team's math and the model accuracy is definitely high. In fact, Jit says it is almost *too* high. He warned me about AI that seems too good to be true.

"I get it," I said to Jit. "I can see that to get to the bottom of this, I'll have to ask a lot more questions and find out what they are using as predictors of churn."

Jit said, "It goes beyond predictors. You can't assume anything. You have to ask them how they defined the business problem itself. For example, what is their definition of customer churn?"

Sometimes I wish Jit would just give me the answer instead of telling me stories and asking me questions. This Socratic method stuff is really frustrating.

"The definition of churn seems kind of obvious," I said. "If a customer stops doing business with the firm, they have churned, right? But I will ask the team about their definition."

His next question was: "Why are they analyzing churn in the first place?"

Again that seems obvious, but I will ask. His next questions were more interesting to me: What are the key drivers of the predictions? What exactly has the AI learned? In essence he is getting at the same question I was going to focus on: What are the key predictors of churn?

It turns out that ManuCo's definition of churn is not so obvious after all. Manufacturing customers don't really cancel their accounts; they just stop ordering.

In other words, for ManuCo, churn looks like changing hairdressers (going to a new stylist and never coming back to the old one), not switching cell phone providers. Churn itself is not an event; it's a pattern made up of a lack of events.

In order to accommodate this customer behavior, the data scientists defined churn as a customer not doing any business with ManuCo for a period of six months.

Why six months? I asked. The data scientists had tried a few different timeframes (one month, three months, and six months). The model for predicting churn on the six-month horizon was most accurate, so they went with that definition.

The way the team made that decision spurred me to dig deeper with their internal 'customers': the business users who were unhappy with the model.

When I asked the business users about the definition of churn and why they were analyzing it to begin with, the real source of the problem started to emerge.

Their definition of churn—and their motivations around why they cared about it—differed substantially from the definition the data scientists settled on. Jit keeps hammering home the need for data scientists to collaborate closely with business users and ask questions. It seemed clear that the

data scientists built this model on their own, without taking their internal 'customers' needs into account. That explains a lot.

The business users' approach toward churn was far more nuanced. They wanted to identify at-risk customers and intervene *before* they churned. They knew what that intervention should look like: assigning a dedicated customer success expert, a form of white glove treatment, to those at-risk customers. From their experience, if a customer has not done any business with them for six months, it would be extremely difficult to win them back, even with white glove treatment.

The business users also complained that the model kept predicting that small customers, like startups, would churn. Providing startups with white glove treatment didn't make sense. The cost of retaining them would be higher than their customer lifetime value.

Business users are already skeptical that AI can help them. Handing over a model that doesn't meet their needs made this much worse. This is one of the biggest hurdles in getting value from AI, and it comes up again and again in different settings. In college I would have laughed if you told me the biggest roadblock to AI adoption was what people thought about AI. Now I know better.

The data science team had not communicated enough with the business users to really understand their concerns. That was the crux of the problem.

Jit stresses the importance of communication versus just looking at data and staring at screens. This is still new to me. At school, I became used to a handful of like-minded researchers, with a well-defined hypothesis approved by the primary researcher, not a cross-section of teams at cross-purposes who don't agree on what appears to be basic definitions.

While sometimes Jit's questions seemed obvious to me, the more people I talked to, the more I realized that I couldn't take anything for granted when it came to definitions in business.

When we examined what the overly accurate model was actually based on, the problem was painfully obvious. The data scientists had retained all three definitions of churn in their dataset. Based on this, the AI had accurately learned that if the customer had not done any business with ManuCo in three months, they would very likely not do any business in six months and would thus have a high probability of churn. This was an almost useless insight and it explained why the predictive model was so accurate—and ineffective.

To add insult to injury, one of the key things the AI had learned was that startups were very likely to churn. It was not accidentally flagging startups; it was actually proactively flagging the kinds of companies business users were not interested in.

The solution was fairly simple. First, we pointed the AI at larger customers by removing startups from the dataset.

Second, we set it to look for customers who slowed down their rate of doing business with ManuCo. If a customer typically does $1M of business a month with ManuCo, and their order volume drops significantly after adjusting for seasonality, we want to proactively engage with them to avoid churn.

The real goal wasn't predicting churn; it was *avoiding* churn. Providing white glove treatment to high-value accounts that had reduced their order rate was more effective than fighting over the perfect definition of churn. Of course, our model was much less "accurate" than the original model—but it was much more useful.

And, to be honest, that's something that no data science course ever taught me.

JIT'S TAKE

What Vera experienced here is something I've seen time and again. Data scientists go and build models without ever communicating with business users to define basic terms and business objectives.

An accurate model is not the goal; preventing customer churn by predicting it in time to intervene is the business goal. That's what made the "less accurate" model useful: it suited its intended purpose.

And *usefulness* is really what you're after, as Vera learned during this project. Many data scientists never learn this

lesson and most data science books and software perpetuate this problem by focusing excessively on things like R-square and lift.

These are measures of model accuracy, not business benefit. And if the business does not benefit, model accuracy simply doesn't matter. For example, we may increase the lift of a marketing model such that instead of 10% of users clicking a link on a marketing email, 20% of users now click on it. This seems like quite an improvement! However, if I achieved this by being very selective about which customers I send the email to, my total number of customers who click the link may be much lower in the scenario with the higher lift. Given that my business benefits are tied to the total number of customers I actually acquire, the AI with the higher lift can actually be much worse in terms of business benefits.

Even worse is the confirmation bias that this type of incident creates. Many business users are skeptical of AI. An accurate but useless model that doesn't move the needle is enough to kill an AI project.

Myth

Training AI is all about improving accuracy.

Reality

If you don't fully understand the business objectives, you can very accurately pursue the wrong goals.

Chapter 6

Why Most AI Are Accurately Wrong

In which Vera learns that most AI are trained to optimize the wrong metric.

Today Jit and I have been called in to do an independent audit of the models that data scientists have created at NoBlemish, a multinational skincare and cosmetics firm. Ed Gibson, their VP of Sales, set up the meeting and invited key business stakeholders as well as the whole data science team.

Ironically enough, Ed heard about us from Susan Nathan from ChemCo. She told Ed that she would have spent more money with us to get more accuracy, but Jit stopped her since it wouldn't pay off. I really didn't think the ChemCo project would lead to more work for us!

Ed is convinced that the data science team—and their models, which they are quite secretive about—are not grounded in business reality.

After introducing us, Ed kicked off the meeting. It felt adversarial right from the start. Diane Kim and her team of data scientists looked defensive.

"Here's the situation as I see it. Diane, you and your team have created multiple AI models for predicting the quantity of each product that will be sold at our partner retail outlets each week," said Ed.

"My concern is that you're only sharing one of those models. I don't know what factors the models take into account or how you're selecting the one you've shown me," he continued.

"I respect your team and their work, Diane, and I don't have any data science background. But I do know sales and I'd like my team to work more closely with you to understand what's going on, even at a high level. That's why I brought in Jit and Vera."

I started the discussion. "So Ed, you want to see predictions from multiple models?"

"Sure. Even the meteorologists on TV share multiple models and explain the implications for the forecasts of major storms," said Ed.

"Diane, how do you select which model you show Ed and his group? How do you evaluate your models so you can choose the best one?" I asked.

Diane looked a bit confused then asked, "Do you mean which measures of accuracy we use, such as Precision, Recall, and Log Loss?"

This was definitely data science-speak, and I saw the business

execs start to glaze over.

I said, “Are you selecting the model that is most accurate?”

“In short, yes,” said Diane. “Accuracy measures whether what we predicted actually happened. Essentially a one unit overestimate has as big of an impact on our accuracy measures as a one unit underestimate because we are only considering the extent to which we missed the actual sales quantity.”

Ed jumped in. “Plus or minus one unit doesn’t sound like a big thing, but it has real world implications. Plus or minus a minute is the difference between getting on the train and waiting two hours for the next one. ”

Jit looked at Diane and asked, “To Ed’s point, are the consequences of overestimating demand the same as underestimating demand? What happens in each of those cases?”

Diane replied, “If we underestimate demand, it might lead to a stockout where our product is sold out. Most often the customer simply buys someone else’s brand. We ran a Customer Lifetime Value analysis last year where we found that if a customer tries a different brand, 20% of the time they switch to that brand.”

Ed agreed. “If we underestimate supply, we risk losing customers. If we overestimate, we just store a little extra inventory. Underestimates are definitely a bigger problem for us,” he said.

Someone from Diane's team added, "But there are ways to mitigate that issue, right? Let's say we decide that whenever a customer can't buy a NoBlemish product because it is sold out, they receive a $50 coupon that expires in 2 weeks. If that coupon addresses the concern about the customer buying an alternate brand, the cost of an underestimate would be a fixed $50."

"Taking such a measure changes the nature of the cost-benefit imbalance, but there is still an imbalance to consider. Whether you want to mitigate the risk of a customer switching to another brand is primarily a question of business strategy. Of course the data science team can easily quantify the expected impact of such a change. But either way, the AI needs to consider that cost of a stockout here is higher than a cost of carrying excess inventory," Jit concluded.

Jayson from Diane's team challenged Jit. "I see your point, but while we train the model on accuracy metrics, we do consider such differential cost factors after the fact when we evaluate which model was better."

The look on Diane's face was priceless. It was clear that her team did not always, if ever, conduct such evaluations.

Jit rolled right past that and said, "Of course, the best data scientists do such analyses after the fact, and I am happy to see that you do so as well. But, as a data scientist, what would you say if I told you that I optimized my algorithm on one metric and evaluated it on another?"

Jayson responded sheepishly, “Of course. It’s a cardinal principle of training AI that it should be focused exactly on what you want to optimize as opposed to a proxy.”

“Exactly right,” said Jit. “And in the same way, if we want to maximize the financial benefit of the AI, then we should simply train it to optimize the financial benefit instead of first training it to optimize accuracy and then evaluating it based on the expected business impact.”

“Why choose a model for accuracy and then test it for ROI? This is like shortlisting marathon runners by first making them run a 100 meter sprint. The best marathon runner may have already been eliminated at the end of the sprint and never got to be evaluated based on their marathon skills,” Jit continued.

Then one executive piped up, “I see your point when it comes to sales. But, I am in operations and our problems don’t easily translate to the world of costs and ROI. What if I can’t come up with the costs of the different types of errors?”

Jit was in the midst of taking a sip of water so I stepped in to answer this question.

“The costs of errors don’t need to translate directly to dollars, though they often can be translated to dollars. For example, in an operations use case such as preventive maintenance on your delivery trucks, we might look at how long it would take to handle a false positive versus a false negative. So, if we unnecessarily replace a part, that takes 15 extra minutes

during maintenance, but if we fail to replace a part, the truck might break down, delaying the shipment by 3 hours. In this case our business impact measure may be in terms of hours as opposed to dollars," I explained.

Jit chimed in to clarify. "Maybe I should explain what a false positive is. Think of it as a case where the AI predicts something to be true and it is wrong. So, it predicts that a part will fail and it actually does not. A false negative is the opposite where it predicts a part will not fail and it does."

"These kinds of prediction errors will always crop up in the real world. In some cases, it may even be a question of relative risk incurred by the business in the case of a false positive as opposed to a false negative. But, as long as the expected impact of the two kinds of errors is not exactly identical, then we should train our AI on net impact as opposed to simple accuracy," he argued.

"Even in cases where we don't precisely know the costs of the two different types of errors, as long as we know they are not identical, it is better to use our 'best guess' of the costs because if we don't guess, then we are simply saying the costs are identical—which we know from experience is wrong."

Diane quickly summarized the discussion for her colleagues and said, "Jit, you are essentially saying that every AI I ever trained was trained on the wrong metric. I am not sure I am fully convinced, but I will think about that. If you are right, I need to think deeply about how to calculate the net impact

of the AI and then we should train the AI to optimize that metric directly."

How to calculate net impact of AI

You have to calculate net impact slightly differently depending on whether you are predicting a binary outcome such as Win/Loss or Infected/Not Infected as opposed to a continuous variable such as quantity sold or length of stay at a hospital.

Let's tackle the binary case first using an enterprise sales example. If the AI incorrectly says that a prospect will buy (false positive or overestimate) then my sales team may waste $500 making unnecessary calls trying to sell to a customer who was never actually going to buy. However, if the AI incorrectly tells me that a prospect will not buy (false negative or underestimate), then my sales team may miss out on a $100,000 deal. I may be willing to work almost 200 unnecessary deals at $500 each to make sure I do not lose that $100,000 deal.

There are four possible scenarios:

1. True Positive: The AI predicted success and was correct. The average benefit is a $100,000 deal.
2. True Negative: The AI predicted failure and was correct. This means we do not spend an average of $100 pursuing this deal unnecessarily.

3. False Positive: The AI incorrectly predicted success. As a result our team tries hard to close the deal and typically spends $500 before realizing the deal won't close after all.
4. False Negative: The AI incorrectly predicted failure and we thus missed out on pursuing a $100,000 deal.

When we evaluate an AI after training it, we calculate something called a Confusion Matrix that essentially tells us the expected proportion of each of these four kinds of outcomes. Let us say that for a specific AI, out of every 100 predictions we expect 25 True Positives, 50 True Negatives, 10 False Positives, and 15 False Negatives. The net benefit in this case is $1,000,000 as calculated below.

Translating a Confusion Matrix into Business Impact

Category	Count	Benefit & Cost	Total Impact
True Positive	25	$100,000	$2,500,000
True Negative	50	$100	$5000
False Positive	10	-$500	-$5000
False Negative	15	-$100,000	-$1,500,000
Total	100		$1,000,000

A different AI would have the same relative benefits and costs but a very different confusion matrix. Thus, each model would have a very different net benefit. Note that in this case the cost of a False Negative is so high relative to a False Positive that an aggressive AI would do better even though it would have lower traditional accuracy ratings than a more conservative AI. Essentially, the financially better AI would be willing to make almost 200 False Positive mistakes to avoid making one False Negative mistake.

We should also consider that certain organizations may want to only focus on hard costs as opposed to soft benefits such as costs avoided (True Negative) or opportunities not pursued (False Negative). Such organizations would only look at True Positive and False Positive benefits and costs.

For continuous variables, like quantity sold in the NoBlemish use case, the math is a little more complicated but can be handled similarly. First we need to figure out whether the consequences of errors are proportional to the amount of overestimate/underestimate or not.

In the NoBlemish case, if they overestimate the quantity, they incur inventory-carrying costs. For example, they may incur a daily inventory carrying cost of $0.01 per unit not sold. The cost of an underestimate would translate into lost revenue from people who buy a different product just once, delayed revenue from people who wait and then buy a NoBlemish product after all, and potentially lost Customer Lifetime Value from the 20% of people who switch permanently to an alternate brand. A strategy like the coupon mentioned earlier might mitigate this.

Estimating this is complex, so we might decide to only include some of these costs, such as just the per unit lost revenue from the people who buy an alternate brand this one time. The math does not have to be perfect; it just needs to be better than saying overestimates and underestimates have the same impact, when they clearly do not.

Once we figure out the costs, the math to evaluate an AI is similar to what we did before. We just test each AI to determine how it will perform as regards estimating the quantity accurately, underestimating, and overestimating, and then calculate the resultant net benefit. When the AI is trained, we let it optimize for this net benefit instead of just accuracy.

"But I still have a question, both for you and for the business team," said Diane. "And by the way, Ed, I'm glad to bring you into the process with us so we can learn from each other about how to make these models better. The more eyeballs on the models, the better." The data science team nodded.

"I still have a problem with showing results from multiple models, though," she said.

"It's one thing to meet and work on this together; it's another for an AI to show multiple results to salespeople or sales assistants," said Diane.

Jit continued, "First let's talk about what you should do as a Data Science team, then let's talk about how to frame it for the end users."

"As a Data Science team, you should absolutely run multiple models to select the best model for your use case and continue running multiple models in parallel if it makes business sense. In fact, even in cases where you have a favorite model that you have extensively tested and tuned, you should always be working on the next model that might eventually become better than your current favorite. Every model can be improved and every model can go out of tune. Running multiple models is simply a best practice," said Jit.

"We often run multiple models in parallel for the same problem to see which model is doing best," explained Diane. "Sometimes one model does better one week but worse the next."

"But coming back to end users, if we keep switching models week by week that can confuse users who may notice subtle differences. At the same time, if we show multiple predictions, we are concerned that user adoption will suffer if the predictions are inconsistent," said Diane. "We know that any AI can generate a bad prediction once in while, and that any two models will differ on certain transactions, but most users expect AI to be accurate."

Getting to the deeper question, she asked, "Will revealing the fact that AI is not perfectly accurate all the time affect broader adoption?" The business execs nodded, sharing her concern.

"Whether you should reveal the results from multiple models to the end user is more of a business question," said Jit, turning to the business executives in the room.

"Before we dive into that, does everyone understand that two different models can have the exact same overall predictive accuracy but give completely different predictions for the same scenario?" asked Jit.

Some of the executives looked a little confused, so Jit explained. "Predictive accuracy is a measure of how often a model gets the prediction right versus wrong. It is calculated across a sample set of transactions and represents how accurate the model will be for other similar transactions."

"Let's say we have two AIs that each detect the color of different objects," said Jit. "One of them always gets the color red wrong while the other always gets the color blue wrong. If the test sample had one green object, one red object, and one blue object, both AIs would be about 67% accurate. But if we looked at a specific object that happened to be red, only one of the models would be accurate."

"If an end user sees that one AI thinks the color is red but another thinks it is blue, they may start distrusting both AIs because all they see is that the AIs are inconsistent."

One of the executives muttered, "Well, we should only show the correct prediction, right?"

Jit responded, "I wish it were that easy. At the time of making the prediction, we actually don't know which AI is correct. All we know is that these AIs perform equally well on average across a set of test transactions. We don't yet

know the 'correct answer' corresponding to the prediction we just made."

"Of course in the real world, at some point we can always see what actually happened, the correct answer if you will, and then go back and reevaluate the accuracy of each model based on this new information. If over time one model turns out to be more accurate, we would shift to using that model preferentially," said Jit.

Diane asked, "But Jit, you haven't told us what the best practice is. Should we show both predictions or just choose one?"

"I hate to say this," Jit responded, "but it really depends. If the models disagree all the time, you may have a real trust issue on your hands. In such cases, you would probably have to choose one of the two models while continuing to test the other one."

"However, if the two models agree most of the time, and only disagree rarely, you can try showing the users the extent to which the two models disagree."

"There is nothing wrong with users knowing that AI is not magic. In fact, you might even want to collect user feedback on which prediction they found more likely. If users consistently side with one of the algorithms, that can also be an useful input into your model tuning activities."

"I see," said Diane. "I'll need to get more guidance from

each of you to decide how we frame this for your users." Ed finally smiled.

For a meeting that started out confrontational, it ended on a high note.

JIT'S TAKE

When we evaluate AI, we need to always keep business realities in mind. In every economics class I ever took, the professor started with 'Let's assume there are no taxes.' Such approaches are fine in academic settings but in the real world we need to remember that there are certainties like death and taxes.

Modern AI is undergoing a fundamental shift from academia where complexities like relative costs of errors can be assumed away to make the math simpler, to the world of business where overestimates and underestimates have tangible and very different consequences.

If you think the benefit of a correct prediction is exactly equal to the consequences of an incorrect prediction, then you are training and evaluating AI based on academic as opposed to realistic metrics. In business, things are rarely symmetrical.

Also remember, when it comes to the adoption of AI, user trust is crucial. If the user does not trust the predictions or recommendations, they will simply disregard that information and will not benefit from the AI.

But does giving users the false impression that AI systems are infallible really foster trust? Let's treat users with respect and make it transparent to them that this powerful technology has limitations. I fully believe users will reward us with their trust. If we present AI as magic, we will inevitably breach their trust and that is very difficult to recover from.

Myth

It's ok to assume away much of the complexity of the real world when we train AI.

Reality

You can't train AI in the theoretical world where costs and benefits are the same and hope to have a realistic outcome.

Chapter 7

One Model or Many?

In which Jit and Vera convince the client that iterating is the key to deciding how many models you really need.

It turned out that the Data Science audit at NoBlemish was actually more of an audition. NoBlemish has created models across sales, marketing, and operations use cases including optimizing inventory levels and predicting product defects. However, their teams can't agree on whether they should have different models for each country they operate in or consistent models across all countries. Diane, the Head of Data Science, brought us back to help resolve this issue.

Naveen Madhav, the country manager for India, was one of the executives asking for country-specific data science teams and AI models.

Naveen said, "I looked at the models developed at headquarters. For example, the size of the bottle is an important factor in predicting which promotion will be most effective. Well, most of my sales in India are in the form of single-use packets, not large bottles."

"The retail chain is another really important factor in the AI model, but the bulk of my sales happen through mom and

pop stores, not major retail chains," he went on. "How can this model ever be useful in India? I can get data scientists relatively easily in India. It makes sense to have my own team and my own AI."

His concerns seemed well founded, but Bob, the Global CIO, made the exact opposite argument, which also had merit. "Sure, in some countries, we sell bottles and in others we sell single-use packets," he said. "Those are just different form factors for the same product. There is no reason why that should stump the AI."

"But fragmentation is a mistake I don't want to repeat. Decades ago we made the mistake of setting up completely different ERP systems in each country," Bob continued. "We are still trying to undo that mistake and get to consistent ERP systems. We should not repeat the same mistakes as we start deploying AI," said Bob.

Diane chimed in with resource constraint concerns. "Given that we do business in 50 countries, it might simply be impossible to create custom models for every country. In fact, if we need custom models for each country, why not custom models for each brand or for each distribution center?"

I took note of these different perspectives and went off to conduct my research.

I quickly figured out that the country manager for India was correct. For some of the countries, the sales patterns were so fundamentally different that the accuracy of the AI increased

if we trained separate models for them. However, there really was no need to have different models for each country.

There were really three groups of countries that behaved very similarly within the group but differed quite a bit from countries in other groups. But we also found that this pattern was not specific to countries. For example, NoBlemish recently introduced a device to be used with certain NoBlemish products. This device was significantly more expensive than anything else NoBlemish sells.

While the AI models were good at predicting sales of other NoBlemish cosmetics, it turned out that the AI did much worse when trying to predict the sales of this device.

In fact, when I removed the data related to this device from the dataset before training the AI, the resulting model was even better at predicting the sales of the cosmetics. In this case, building two separate models increased the overall accuracy.

As Jit and I walked the NoBlemish executives through our findings, Diane asked, "Are you saying that this is a resourcing issue? And that we should do many small models instead of one big one as long as we have the resources to create such models?"

Jit responded, "Well, you will find that you hit a point of diminishing returns. If you had created a separate model for each country, you would have achieved just slightly higher accuracy compared to creating three separate models for the three groups of countries."

"In fact, it is even possible that for some of the smaller countries, the model accuracy would drop slightly because there would be insufficient data for the country-specific AI to learn certain useful patterns that it could have learned from the broader dataset," explained Jit. "In such a case, even if you had data science resources available, I would not waste their time on creating additional models."

Diane looked serious. "So how do we decide whether to create three models or fifty?" she asked. "Wouldn't I need to create fifty models to determine that they were barely better than three models?"

Jit responded, "In a perfect world with infinite resources, that might very well be what we would do. In the real world, I have seen a couple of approaches work well. Let me share the stories of two different clients who approached this problem in different but equally effective ways."

"The first client's AI initiative was led by an experienced and pragmatic data scientist," said Jit. "He first created an AI model and looked at the most important predictive variables in the data. For example, if the country variable was a good predictor, he would look at the data from each country to detect clusters of countries that behaved similarly. Then he split the data into those clusters, retrained new models for each cluster, and checked whether the overall accuracy improved. He continued this approach with each of these smaller models until he ran out of time."

"This is obviously not a perfect solution," Jit admitted. It specifically runs into problems where combinations of variables are more important predictors than individual variables. However, it is a fairly valid approach to solving this problem."

"The second customer's AI initiative was led by an experienced and curious VP of IT," said Jit.

"He was not very comfortable with the idea of evaluating different predictive variables, but he was an expert at monitoring complex systems. He set up a process for checking the accuracy of each prediction and looked for underlying patterns to any problems," continued Jit.

"For example, he noticed that the AI was much worse at predicting sales when it involved a specific distributor. It turned out that this distributor only sold to government customers while every other distributor sold to commercial customers. Because government purchasing processes are significantly different, he created a separate model for government purchases," said Jit.

"By monitoring the AI closely and splitting off specialized models for cases where the AI was least accurate, he quickly built up a very powerful stratified model," said Jit.

"Over a span of two months, he found that he had tested more than 200 models. But because each time he was focused on a specific pattern of low-quality predictions, he was able to systematically navigate the process instead of getting

overwhelmed. If he had started out saying he would evaluate 200 models and then decide on the best combinations thereof, he wouldn't have known how to even start deciding which 200 models he should create," Jit concluded.

I chimed in at this point, explaining that these two approaches were of course not the only ones. For example, some customers have tried using hackathons to create and compare many models in a focused way, while others solicit feedback from users to determine cases where a specialized model would be more valuable.

What these approaches have in common is that they are systematic ways to improve the AI iteratively and quantify those improvements.

JIT'S TAKE

I have always believed that all analysis is iterative. You look at what you have and find ways to make it better. AI model training follows exactly the same pattern. Your first model will rarely be the best model you can come up with, and every model can be improved a little bit.

Based on the unique skills of your team or the culture of your organization, craft a systematic approach to iteratively improving the models. The more you incorporate objective observation and user feedback into deciding where to focus your energies, the easier it will be to manage this iterative process.

Over time the number of models you deploy will grow, but as long as they are increasing overall effectiveness, that is not a matter of concern. At the same time, don't get forced into creating multiple models by organizational politics. That leads to unnecessary fragmentation of models and may actually reduce model accuracy because the AI is learning from a smaller dataset.

Myth

You are looking for a single best AI.

Reality

You will almost always end up with a few specialized AIs that work in parallel.

Chapter 8

Why Pilots Don't Work for AI

In which Vera and Jit explain that customers who are happy with an AI pilot may not be happy for long.

Today, Jit and I are advising SoftCo, a major software vendor, on their new AI product. Their product, InvoiceClassifier, predicts which invoices will be paid in a timely manner and which invoices will be disputed.

Aisha Marks, the product manager, summed up their concerns. "We had a great set of pilot projects where there was no evidence of a problem. But just two months after launch, customers claim the product is not working as they expected."

She looked at me, "We know that nothing has changed in the product, so what is going on?"

"Out of curiosity, what was the definition of success for the pilot?" I asked.

"Well," said Aisha, "we had to have at least 50 customers

accept the predictive models and agree to deploy the AI in their companies," she explained.

"And, on what basis did they 'accept' the predictions?" asked Jit.

"Based on some sample data, we predicted which invoices would be disputed and the customers confirmed that the predictions made sense based on their past experience," explained Aisha.

A red flag went up immediately. I suspected another case of black-box AI. To confirm my hypothesis, I asked, "Did you provide any explanation of the basis for the AI's prediction?"

Aisha was a bit defensive. "Our product explains exactly what drove each prediction, and these drivers made sense to the customers. For example, our customers might know that certain large companies pay late but always pay, while certain smaller companies often had payment problems. In the InvoiceClassifier predictions, you could see that it predicted a higher probability of payment disputes when the end customer was a smaller company."

Okay, so maybe black-box AI wasn't the problem here. Because InvoiceClassifier explains the reasons behind each prediction, and the users accepted the explanations as well as the predictions, there should not be a problem, right? If the AI learned the wrong thing, the user would just need to look at the reason behind the predictions and see where it had gone wrong.

The best way to figure this out will be for Jit and me to interview a bunch of customers and see what we can learn.

As we started reviewing the data and started speaking with end users, it became clear that while InvoiceClassifier provided some explanations about its model, it did not provide sufficient information so that the user could really understand what was going on.

And as we kept speaking with customers, the list of problems with InvoiceClassifier pilots kept piling up. Some customers had piloted the product in a specific geography and then deployed it worldwide even though the AI had not been trained on global data.

A few customers had become victims of their own success where their focus on collecting certain types of invoices had impacted the payment behavior of the corresponding customers and thus the original model went out of tune. The conversation with Aisha was going to be a very difficult one.

Once we walked Aisha through all of the customer evidence that showed why her successful pilots had not translated to successful deployments, she asked, "Are you saying we should shut down this product because it won't work?"

Jit responded, "There is definite evidence that InvoiceClassifier works during the pilot. The question is how we can help you change your pilots in such a way that you focus on making customers successful on an ongoing basis instead of just having successful pilots."

"What exactly do you mean by that?" Aisha demanded.

"Have you ever heard of white hat hackers or penetration testing?" Jit said.

Aisha was thrown off guard. "They're used in testing security, right?"

"Testing AI is more like penetration testing than you would think," said Jit. "The people who work on your pilots should essentially look for ways to break the AI, looking for potential seeds of future problems," said Jit.

I added my perspective. "When we interviewed your deployment teams, it was clear that they saw their job as making the customer happy with the pilot, which is understandable."

"If we exclusively focus on making the pilot customer happy in the short term, we would obviously be better off not looking around for potential problems. Why create problems where none exist yet?" I continued.

"If, on the other hand, your goal is to make customers successful in production," added Jit, "we will do our best to find ways to break your pilot AI so we can better anticipate and prevent problems when we deploy the AI in production."

"Would you like us to work with one of your customers and take more of a penetration testing approach?" I asked. "We'll be glad to train your pilot team on what to look for."

"I'll have to think about which client would be best-suited for that mission," Aisha said. "What you're saying runs counter to all my experience in launching software products. I need to talk to my executive team and get back to you."

"Before you have that talk," said Jit, "There's something else you should know."

"This is something you will need to do periodically. Get teams of experts to do nothing but try to break the AI by testing it on data it has not seen. If the AI passes the tests, then deploy it, but even then your work is not done. You need to stay vigilant and evaluate how well the AI is doing over time. If it starts doing worse in one region or for one product type, retrain and update the AI," he said.

"Listen: I'm used to changing software and then having those changes sometimes break other things," said Aisha. "You're saying that larger exposure can break an AI, even though, as in our case, the code base is intact."

I gave Aisha a sympathetic but serious look. "The difference with AI is that it is not just your code but also the data it is trained on that determines the final product. Think of normal software as a baked cake that you might customize by using different decorations and maybe by changing its shape. An AI product is merely a recipe and a cake form. The ingredients are the customer's data. Just because the recipe worked with chicken eggs may not mean it will work with ostrich eggs. AI interacts with data dynamically, learns from it and changes in response to it."

JIT'S TAKE

In enterprise software, the pilot is a well-respected rite of passage. How do you determine whether a piece of software works as promised?

Of course, you pilot it in one department or one process and if it works, you roll it out across other parts of the company. The implicit assumption here is that if the software worked in one context, it will work in another.

AI does not work this way. If I trained the AI on data from the US, it may or may not work well when we apply it in China. If we trained it on data from the first half of the year, it may not work well on data from the second half of the year.

As the market evolves or our business evolves, the AI may not stay in tune. The pilot paradigm simply does not work for AI.

Think about how to break the AI instead, as penetration testers do to vet the security of networks.

Myth

If it works in the pilot, it will work when you roll it out.

Reality

AI learns based on the data it was trained on. Unless your pilot data was a perfect reflection of the overall data, you will always have to retrain the AI as you roll it out more broadly.

Chapter 9

Predictions are Nice, Recommendations are Money

In which Vera learns that for most people, actionable recommendations are better than predictions.

Today I am meeting with Janelle Ray, the CMO of IntFashion, one of the largest fashion houses in the world. IntFashion is famous for their extensive deployment of AI technologies, so I am very interested in seeing what they want us to help them with.

Janelle started by explaining how widely they are using AI, but every example she cited seemed to focus on predictions of various sorts. They are predicting:

- Numbers, such as how many socks they will sell in a specific store this week
- Probabilities, such as the probability that they will run out of socks to sell this week

- Which ‘class’ something belongs to, such as using image recognition in inventory management where the AI detects the style of a sock that had its inventory tag damaged.

This is not surprising given that one of the most common uses of AI is in making predictions. But for a company using predictions so widely, we were surprised to see no references to using AI to recommend what people should do.

Maybe that’s why I’m here.

After Janelle walked me through the state of the art at IntFashion, she came to the core of the project. She said, “We have had a lot of success using AI inside the company, but now we are trying to get other retailers who sell our brands to use similar predictions. That project has not gone very well. Can you help us bring our partners up to speed in using AI like we do?”

It sounded like a reasonable goal. Given IntFashion was clearly making a lot of money from their use of AI, why wouldn’t their partners want to get on the bandwagon?

Janelle also confirmed that IntFashion was not asking partners to pay for the use of the AI. They just wanted to increase the sales of their products through these partners.

As I started interviewing the partners though, I quickly realized that they did not see IntFashion’s predictions as beneficial.

A store manager said, "IntFashion told me that I would sell out of a certain item, so I removed it from the sale section. Then it did not sell out and I had to deal with excess inventory. I looked really bad in front of my boss."

Shayla, IntFashion's AI lead, and her team were listening in on the partner interviews. After the meeting, during our debrief, one IntFashion analyst said, "But we didn't tell him to stop discounting the product. Why did he do that? Of course if he removed it from the discounts section it would not sell out! The prediction was based on the facts we knew at the time of making the prediction. If he changed something as fundamental as the discounting, then of course the prediction would not be accurate anymore!"

It was a fair point, but I had a very important question that I had wanted to ask since I saw Janelle's presentation, "So what did you expect him to do with the prediction? What was the recommended action?"

Shayla responded, "Well, there are many ways a store manager can react to this kind of a prediction. For example, he can slightly reduce the discount rate, or think through what alternative item he should sell once this item sells out. There are so many ways to address this prediction that we don't want to cramp people's creativity about what they do in response."

I wanted to shout that this particular store manager's decision to not discount the item seemed a perfectly reasonable

creative response, but Shayla is a client and it never helps to fight with clients.

Taking a deep breath, I asked, "Could you perhaps have evaluated these potential responses and predicted the probability of stockouts at different discount levels for example? This seems like a complex optimization problem perfect for an AI."

My question triggered a heated debate among the IntFashion team. The debate continued over dinner and drinks that night.

Eventually, by reading between the lines a bit, I managed to form a fairly clear view of what had actually happened that caused the team to not include recommendations in their AI solutions.

When they trained the original AI, executives were very happy with the quality of the predictions. They accepted that sometimes the predictions were not exactly accurate because people tended to change things between the time the prediction was made and the results were observed. The execs accepted that the predictions would only be approximately accurate as long as they were useful and had business impact.

For example, the prediction at the retailer where the stockout did not happen would have been perfectly acceptable at IntFashion because the conditions under which the predictions had been made were changed. No wonder the prediction was inaccurate! But when the team started trying

to *optimize* instead of just predicting, when the AI started trying to make recommendations, then it was held to a much stricter set of standards.

If a user took a recommended action, such as setting a specific discount rate, management expected the outcome to be exactly as predicted. The team tried their best to explain that recommendations are never perfectly accurate and that even when the user took the recommended action, unrelated factors, such as inclement weather, could still affect the outcome.

The team was able to show the execs that on balance taking the recommendations significantly improved the business outcomes. But all of their rational arguments fell on deaf ears.

Certain senior executives were obsessed with individual examples where they took the recommended actions and did not get the expected outcomes, and they simply refused to use the AI's recommendations.

Given that predictions were relatively uncontroversial, the IntFashion leadership decided to give up on recommendations and rolled out just predictions instead.

As I started interviewing the partners, I knew we had a problem because the partners were clearly asking for recommendations and some estimate of the expected impact of acting on the recommendation.

As one store manager put it, "I have about 50 things to do

each morning before the store opens. You have to convince me that acting on your recommendations should make it to my top 50 or better yet top 10. Don't tell me I will potentially have a stockout on one IntFashion item. So what if it stocked out? Customers might buy something else from another designer brand. Now if you tell me that by discounting an IntFashion dress 10% instead of 20%, I will make an extra $1,000 in sales this week, you have my attention."

For once I was looking forward to placing a problem in Jit's lap. Let him figure out how to either convince the partner retailers to accept predictions or convince IntFashion to start providing recommendations!

Jit seemed way too comfortable with the problem when I briefed him. Perhaps because he had not been at the interviews, he did not fully realize the extent of the disconnect, so I asked him how he planned to approach this problem.

"What is the real problem here?" Jit asked.

"Well, clearly the problem is a disconnect in objectives between the two parties, right?" I replied.

Jit explained, "I see a potential opportunity for creating a significant amount of value for IntFashion. Both the partner retailers and IntFashion have the exact same goal: making the most sales."

"IntFashion stores are not that different from partner stores," continued Jit, "except for the fact that IntFashion has less

control over partner stores. I bet that the problems they are encountering at the partners are also happening in their own stores. It's just that partner store managers are more vocal about those problems."

I knew what he was about to say, so I preempted him and said, "Ok. I will go interview some IntFashion store managers and see what I can find out."

Turned out Jit was right. As we spoke with the IntFashion store managers, their responses ranged from "Because I have done this for a while, I know exactly how to take the right actions based on the predictions, but I can't exactly explain how" to "I have no idea what to do with the predictions. I just get extra stressed when the AI predicts I will have a bad week and relax a bit when the AI predicts a good week. I am not sure that I really act any differently based on the predictions."

The most positive response came from Amy Hu, a known champion of the AI program, who said, "We come up with a list of things to keep an eye on based on the AI predictions. For example, if the AI predicts an item will sell out, we have an alternative in mind and notify the employees on the floor that they should sell the alternative item if the item sells out as predicted," she explained.

"If the AI predicts we will sell a large quantity of a certain dress for example, we think about what accessories would go well with that dress and notify the dressing room staff

about what to cross-sell when they see a customer trying on that dress. The AI predictions are definitely useful if you act on them," said Amy.

"But to be honest, even I must admit that how we respond to the prediction is more of an art than a science," she said. Amy's response seemed reasonable, but I was struck by her comment about her approach being something of an art form.

How could IntFashion scale this approach systematically across the company if it was an art not a science?

Before I started digging into that question, I decided to validate Amy's response with some of the employees at the store she manages.

Their perspective did not align perfectly with Amy's. As one employee explained, "Sure, we get a list of things to look out for, but many of the predictions never happen, so it can become a bit of a distraction. Moreover, we already try to upsell and cross-sell. I'm not sure how much we're doing differently when we are asked to specifically change our selling strategies for a given week."

Even at the store where AI predictions seemed to have succeeded the most, there seemed to be some debate about the true impact of predictions.

It almost seems to me like the IntFashion store managers are paying lip service to the value of the prediction and then

doing pretty much what they used to before the AI came into play. That is the 'art' part of the equation. That might even be a reason for the resistance to recommendations that offer less room to interpret and act upon as they see fit based on their deep knowledge of the space.

But, I was convinced that recommendations would be very valuable for at least junior employees and partners who do not already have years of experience and domain knowledge with IntFashion.

This fit right in with the conversation we were having with the partners. The partners wanted something actionable—a recommendation—not a prediction. They didn't have time to think about predictions and had mixed results when they tried.

It is clearly time for IntFashion to focus on incorporating the executives' domain knowledge into the AI instead, following a process similar to what we did at BigBank[5] and BigRetailer.[6] That way more employees can benefit from the understanding their managers have developed.

Jit and I decided to openly discuss our concerns with Janelle and present our case. We walked her through the source of the disconnect between the partner retailers and IntFashion, we presented data that indicated IntFashion was actually not maximizing the benefits it could get from the proper adoption

[5] Chapter 2.
[6] Chapter 3.

of AI, and we provided her our independent evaluation of the work the IntFashion Data Science team had done that indicated IntFashion could get significant additional benefits if it used AI to optimize processes and make specific recommendations instead of just predictions.

Janelle responded, "I completely get what you are saying, but I have a political problem here. I can't just revisit the decision that was taken by the entire executive team to focus on predictions, not recommendations."

Well, we had a solution for her. There was a problem with IntFashion partners not adopting predictions and requesting recommendations. Then we had a separate problem with IntFashion executives not wanting to adopt recommendations. Why not use one problem to solve the other?

Over the next few months we ran a pilot project with a few partner retailers where we provided specific recommendations or action plans for what they could do to increase their sales of IntFashion items.

We made it very clear to IntFashion executives that this was only for external partners; we were not focusing on how IntFashion used AI in their own stores.

But, of course we could not be sure the partners were accurately acting upon the recommendations because IntFashion didn't have the final say in those partner stores.

To allow us to test the recommendations in an environment

where IntFashion had complete control and could enforce the recommendations, IntFashion executives allowed us to run the same pilot at three IntFashion stores as well. The results were stunning.

The partner retailers saw a significant increase in their sales of IntFashion items relative to what they were seeing for other items. Most of the partners agreed to adopt the IntFashion recommendations.

We had clearly succeeded in the main project we had been brought in for. Jit and I were however much more proud of the unstated bonus project we were working on. As we expected, the three IntFashion stores that we were using to test and calibrate the recommendation algorithms also ended up increasing their sales by over 30% relative to other IntFashion stores that were just using predictions.

JIT'S TAKE

The main reason we deploy AI is to improve business processes. While predictions can help with that, specific recommendations make AI more actionable. If we don't act upon what we learn from the AI, how do we actually effect change? The easier we can make it for the users to act on the insights generated by the AI, the more likely they are to act.

People with different levels of expertise benefit from different approaches. AI projects are usually evaluated by senior people, but AI primarily benefits the bulk of your employees who don't have years of experience at your firm.

C-suite executives at many of my clients say, “I already knew what your AI learned. So what is the big deal?”

My answer? I ask them point blank whether their newest employee knows what the AI learned.

Do you want your most recent hire to benefit from the knowledge that your top employees have? Then have AI trained on the work done by the most experienced employees advise the rest of your employees. Think of AI as a way to encode and disseminate institutional knowledge instead of leaving that knowledge locked inside the heads of your top employees.

By the same token, recommendations work well where predictions alone fail. If responding to the AI is an art, it will be especially difficult to train people on it. You want AI to give specific, direct recommendations precisely so you don’t have to train lots of people on how to respond to the AI in complex ways. You just want them to follow simple context-specific recommendations.

Predictions alone are like a system that tells us we will be 30 minutes late for a meeting we are driving to. This is somewhat useful because at least we can inform the person we are planning to meet and apologize for running late.

What we need is a GPS system with real-time traffic alerts that will tell us how to avoid traffic and make it to the meeting on time.

Myth

AI is best for predictions.

Reality

For most people, recommendations are better than predictions alone.

Section III

Major Factors to Consider When Implementing AI

This section is all about concepts that may seem tactical at first, but they are crucial to your success as you adopt AI. In each case it might seem like this is a topic better left to a data scientist or an analyst, but pay attention: these concepts can make or break an AI project.

In the last section we talked about the importance of ROI. Now, can we just delegate the actual calculation of ROI to an analyst? Not really. Business leaders need to be directly involved to make sure ROI is calculated in a way that reflects real-world business objectives and to ensure ROI is monitored over time. Moreover, don't forget a related point from the previous section: manual processes have value too. Thus, the ROI of your AI needs to be higher than the ROI of the manual process it is replacing. Otherwise, you may deploy a highly accurate AI and promptly start losing money because of it.

After you deploy your AI, you can expect it to quickly start degrading. In fact, the more effective your AI is, the more it changes human behavior or market conditions, the more quickly it goes out of tune. If your AI degrades over time, that is not your fault. But if you don't notice that your AI has degraded or don't have a process for continuously retuning your AI, then you have not properly set up your AI project.

You can't wait for clean data. Unfortunately, there is no such thing. In AI, you typically have a choice between clean stale data or dirty fresh data. That does not mean you should not strive for clean data. But you can only really clean data after

you have fully experienced its subtleties. Get started, create an AI, observe what it learns and how it predicts, and stay curious. You will see problems and opportunities emerge. Next, improve your data and iterate. By the way, the day you are sure you finally have your data pipeline perfectly set up, something will change in your data or a new pattern will emerge and you will have to iterate again. If you don't set up your AI project with these realities in mind, you will fail. There are two ways that can happen; you can fall into analysis paralysis by failing to deliver value because you're constantly looking for clean data, or your AI will not be able to adapt to the production data because it differs significantly from the clean data you trained your AI on.

Finally we will talk about data leakage. One way of explaining it is that we are trying to avoid accidentally training the AI with information it would not have access to at the time of making a prediction. Can you think of a case where an AI was deployed with great fanfare based on how accurate it was while being trained, and then the project fizzled within weeks or months of deployment? More often than not, data leakage was a factor in why the project failed. If you want to succeed in AI, keep an eye on whether or not your data is leaking.

Chapter 10

Quantifying ROI for AI

In which Vera learns why ROI has to be a core component of an AI strategy.

I thought that the project at IntFashion was a major feather in our cap. Using recommendations, we were able to improve the profitability of IntFashion's retail partners significantly. We were permitted to try out recommendations in three IntFashion stores as a reference point. This gave us insight into exactly how recommendations were being handled, and provided a sort of "control group" for the retail partners' use of recommendations.

The three IntFashion stores saw a 30% increase in sales, which I thought spoke for itself.

But then Janelle, their CMO, called us back. She had been nervous about the executives' reaction to recommendations, and she was right.

Some executives were pleased with the results, but others were upset that Janelle had circumvented them by implementing recommendations in IntFashion stores against their express wishes.

Meanwhile, Kevin Hanes, their CFO, had questioned the value of AI in general given that in his view, he had been assured that predictions were creating value and recommendations were not. Now we were asserting that recommendations would create more value than predictions. He was questioning how IntFashion was even defining success and measuring ROI.

We came in the following week. Kevin started off the meeting by asking us, "How do we make sure we can clearly determine the ROI of our AI projects going forward?"

"Well," Jit said, "how did you determine ROI last time?"

Kevin looked at Janelle.

"We determined that our model was about 85% accurate," Janelle said.

"A model might be accurate but it may not actually create value," said Jit. "It sounds like you measured accuracy instead of calculating ROI."

By this point, I had also become concerned about whether IntFashion was actually getting any ROI from their investment in AI. All too often we were hearing that end users were not really changing their behavior on the basis of the AI predictions. If that is the case, there can be no ROI, because in the end business metrics are only affected once business users act upon the predictions to maximize revenue, minimize costs or risks.

How to ensure you get ROI from AI

1. Determine the business net benefits of the process that existed before AI was adopted and ensure your AI's net benefits are higher: All too often AI is sold based on how accurately it predicts the future. But there was a manual process that was in place before the AI, and that manual process also in a sense made predictions about the future. Many AI ROI calculations implicitly assume that the alternative was 'random selection.' Go into any department where AI is being deployed and tell the team that their efforts are currently no better than random, and observe the reaction. Hint: it isn't going to go over well. People and existing processes are often quite accurate. How to determine the ROI? You can use a similar approach to how we calculate net impact of AI in Chapter 6, because manual processes also estimate correctly, overestimate, and underestimate. You can thus calculate the ROI just like you would with an AI.
2. Make sure you have the resources necessary to act upon the AI predictions. The AI has to be designed such that it produces a volume of recommendations that the business users can act upon. If a store manager receives 500 recommendations, she might be so overwhelmed that she would not act upon any of them. At the same time, if we offered just one or two recommendations each day, we might ignore too many useful recommendations.

3. Make sure your users can actually act upon the predictions or recommendations generated by the AI. Recommendations can and should be customized to the user. For example, a store manager may receive recommendations related to which items should be discounted while salespeople might receive recommendations on which items they should upsell or cross sell based on what the customer is buying.

"You're not alone in focusing on accuracy instead of ROI," Jit said sympathetically. "Often AI are evaluated based on how well they would do versus random chance, like a coin toss. But if you want to have a good sense of the expected ROI from rolling out an AI, you should look at the ROI of the way you ran your stores before the AI was deployed and then compare it to the expected ROI of the AI."

"Is ROI calculation really the only way?" said Janelle. "I've heard some consulting firms have specialized surveys they use to evaluate whether users see value in an AI project. Surveys are perfectly valid ways of evaluating whether our users see value, right?" she said, looking at Jit.

Jit said, "I can't recommend them for this purpose. After evaluating several survey approaches, we found that surveys are more useful in evaluating an organization's willingness to try AI than actually evaluating the impact of a specific AI. Users who were excited about using an AI system voted positively on such surveys before and after using the AI. Similarly, people who were significantly resistant to using

AI voted negatively. While the survey results did shift somewhat based on the users' actual experiences in using the AI, the shift was small relative to the impact of their prior attitudes. The surveys didn't translate into quantifiable ROI."

"Given how fast the market's changing, do we really need hard numbers to move forward?" argued Sam, the VP of Sales. "I heard LowEndFashion speak about their use of AI on a podcast and they doubled their sales in a year. We don't have time to wait on this. We're seeing significant results already in those three stores where we rolled out recommendations."

"But," Kevin interrupted, "that's just one story. My friend who is CFO at MajorFashionBrand tells me that they have continued to sink money into AI projects and they are not paying off. That's why I say we need numbers, not stories. Jit, how would you calculate the ROI of an AI project?"

This is one of Jit's favorite topics, and he was happy to walk them through how to calculate the ROI or net benefit of an AI.[7] The key difference is that for an ROI calculation you would typically only include hard benefits and costs, while for a net benefits analysis you would also include soft benefits like cost avoidance and soft costs like missed opportunities.

Kevin said, "OK. Let's do it right this time. I want to see ROI calculations for any AI you are planning to deploy into production. But I am still concerned about one thing. I've heard that an AI's accuracy might degrade over time and we

[7] See "How to calculate net impact of AI" in Chapter 6

would have to retune it periodically. This ROI calculation is a one-time thing. I need to know how we would monitor the ROI over time in a more scientific way."

Janelle said, "What if we do something like A/B testing, like we do in marketing?"

"That is a great idea," said Jit. "Let's talk through how that would work and see if we all agree that this is the best way to proceed."

"We already have recommendations running in three stores. We could set it up so that the AI would calculate all of the recommendations it could come up with, but would randomly not present a certain percentage of the recommendations to the users. The users and even the data science teams would not know that some of the recommendations were being suppressed," said Jit.

"This is very similar to the way A/B testing[8] is conducted. We will then compare the outcomes and ROI for the cases where the recommendations were presented to users, to the base case where the recommendations are generated but never presented. This approach is conceptually no different from comparing the effectiveness of the manual process before AI to the expected accuracy of the AI," he concluded.

[8] https://en.wikipedia.org/wiki/A/B_testing. In A/B testing one group of users get experience A while the other group get experience B and the differences in the results of the two groups is used to evaluate how much more effective approach A was relative to approach B.

The executives accepted our proposal, enabling us to move forward.

We implemented the approach for those three stores, and came back again. While individual outcomes varied quite a bit, there was statistically sound evidence that using the recommendations was leading to much better outcomes.

In fact, we could statistically estimate what would have happened if we had not suppressed some of the recommendations. The difference between what actually happened when users were not provided the recommendations versus what actually happened when they were provided the recommendations, allowed us to quantify the expected financial impact of adopting the new recommendations AI more broadly.

Moreover, we addressed any concerns about user behavior affecting the results because this random experiment was run without any way for the users or even the data science team to become aware of the experiment. Any external factors such as weather events would have equally affected the recommendations that were presented and those that were suppressed and thus did not materially affect our analysis.

In our follow-up presentation with the IntFashion executives, we included a slide with the results of the A/B test. Because the AI was designed to present recommendations that maximized revenue, the slide was presented in terms of the additional amount in dollars each of the stores made across

those two months by adopting the recommendations made by the AI.

For most of the three-hour executive presentation we ended up staying on that one single slide. Some executives spent a lot of time trying to figure out whether we had made any mistakes in our approach.

While we all agreed that there was room for further improvement, and that there is always some uncertainty in statistical approaches, the fundamental approach was eventually accepted by everyone.

Next, they wanted to know what would happen if this AI was rolled out across all IntFashion stores.

Now Jit had to explain to them that we couldn't just use the same AI we created for the three stores to roll out recommendations across all IntFashion stores. There were significant differences between the three pilot stores and the rest of the IntFashion stores. Thus, we needed to create a set of AIs trained on the data from all the IntFashion stores.

But, we could use the same process to evaluate the ROI for each store and then monitor it over time using the A/B test approach. Janelle was happy because she was getting real provable ROI from AI. But, I think Kevin was happiest. He concluded his last meeting with us with, "I had a visceral problem with AI because it seemed like I was being offered magical results but I would have to take it on faith. I have faith in God but all others better bring me clear ROI calculations."

JIT'S TAKE

There are AI vendors that brag about how almost a billion models have been created using their software. However, none of these vendors talk about exactly how many of these models have actually been deployed.

Why wouldn't you brag about that number? My best guess is that less than a million of these models have ever been deployed, which means only 0.1% of the models created by such software were ever deployed. In the rare cases where these models are deployed, they often fail to create value. Why? Because the models were selected based on accuracy instead of ROI or business benefit delivered.

For an industry focused on mathematical techniques at scale, the AI industry has been curiously coy about measuring ROI in scientific ways. Measures like predictive accuracy and graphs like a confusion matrix may indicate how well a model is doing in a theoretical sense, but if AI is going to be deployed at scale at enterprises, we have to objectively measure the ROI of such projects.

Note that not all AI projects should focus on ROI. When you are in the experimentation phase, you should just focus on failing fast and learning fast. However, when you are ready to deploy an AI at scale, you have to objectively calculate the ROI. Even if you can't be as rigorous as in the IntFashion example, try some form of A/B testing to objectively evaluate the benefit of the AI.

Myth

If an AI is really accurate, it creates value.

Reality

You need to translate the AI's accuracy back to the actual business impact. You can't just assume accuracy equals ROI.

Chapter 11

Don’t Set It and Forget It

In which Vera learns that the unintended consequences of an AI gone wrong can easily overwhelm the strategy of the company.

Almost a year after my very first project, Todd from ManuCo called back and he didn’t sound happy.

The model I created for ManuCo in my first project at Foundation Consulting predicted expected profit and probability of success for each sales opportunity. And for the first few months, my solution worked perfectly. Both win rates and profitability went up and everyone was happy. Then six months after go live, there was a problem: profitability started trending down, even though win rates had skyrocketed.

While the ManuCo team was unable to figure out exactly what was happening, they were certain that it was related to my model.

I went to Jit in a panic. My reputation was built on this

project and now the customer was talking about ending our relationship because our "solution" was possibly causing profitability to go down—the very issue we were brought in to address.

I didn't know where to start figuring out what had gone wrong. Thankfully, Jit had a couple of suggestions.

His first suggestion was predictable. (Jit talks about people more than AI.)

"First, look at human behavior. Whenever you change a system, people's behavior changes, sometimes in unexpected ways. This is especially true in sales because of commissions. Salespeople are often laser-focused on maximizing their sales commissions and that can lead to unintended consequences."

"Ok, what else?" I said.

"Look at the data," said Jit. "How has the data changed since you created the model? Has ManuCo entered new markets or introduced new products? Or has the relative importance or behavior of different products or markets changed in the last year?"

Well, that was a start. Todd was adamant that they had not changed any products or entered any new markets, so I decided to track down the human behavior angle first.

Jit had told me a story from his Harvard Business School days. Apparently, a professor in the incentives class was famous for an exercise where students negotiate deals with each other based on incentives known only to the individual student (and the professor).

Given 80 motivated students with different levels of negotiation skills, you would expect widely varying results. Yet the professor was famous for writing down the correct results before negotiations started. He knew their incentives, which enabled him to predict the outcome with uncanny accuracy. Incentives are powerful.

When I asked ManuCo about incentives changes, they explained that in an effort to increase the adoption of the AI model, they had instituted awards for people who successfully sold to prospects rated highly by the AI. This reminded me of another one of Jit's stories.

At a leading bank, mortgage processors were compensated on the number of mortgages they processed in a month.

It turned out that the most highly compensated employees had focused on only the easiest mortgages and allowed more complex ones to lapse. The lapsed mortgages counted against them, but they could process many more easy mortgages in a month, and get paid more if they did not 'waste' time on harder mortgages.

The problem? More complex mortgages were often more profitable for the bank. And if customers lost their chance to

buy their dream house because of the delay, they might leave the bank altogether, costing even more.

I was going to be extra careful about these new incentives ManuCo had rolled out. People work to their incentives and if they are not well designed, it is easy to incentivize the wrong behavior.

Now I had a key question for the ManuCo team. I had created two models: one for the likelihood of winning and one for predicting profitability. Which one were they using for incentives?

Their answer? They wanted to consider both factors and so they simply multiplied the two predictions and created the incentives based on that combined number. This seemed reasonable for a moment, until I realized what they really wanted to focus on was their overall profitability.

To achieve their goal, they had to maximize expected profits for each sale. Let me explain: A $100 deal with 10% profitability and thus a $10 expected profit is much less attractive than a $10,000 deal with 5% profitability and thus an expected profit of $500.

Deal Size	Probability of Close	Profitability	Expected Profit	ManuCo Incentive Metric
$100	5%	10%	$10	5% x 10% = 0.5%
$10,000	8%	5%	$500	8% x 5% = 0.4%

Even though the second deal had a higher expected profit and a higher probability of close, the ManuCo incentive system would have prioritized the first deal.

ManuCo had incentivized winning a larger number of sales with higher profitability versus maximizing their overall profits. It was easy to fix this, but it still did not fully explain their profitability problem. I took a hard look at the other factor Jit had suggested: the data.

I love data. Data doesn't lie (unlike some of the guys I've dated). I compared the original training data with more recent data.

The biggest difference was a surge in deals creating components for other manufacturers to use in their products. These custom projects seemed to have very low profits. Oddly, most custom projects were done for startups. When I had analyzed the data last year, almost every custom project ManuCo took on was for Fortune 500 companies.

ManuCo had clearly changed its business model.

I showed the data to Todd, and asked why he didn't tell me about the decision to focus on selling to startups. Even when faced with clear evidence, Todd denied making any such changes. I left pretty quickly because it was hard to hide my frustration.

When I told Jit that I thought Todd was hiding something from me, I was unprepared for his reaction.

He laughed. "Always assume human error and fallibility before you assume wrongdoing," he said. I looked at him skeptically and decided to dig deeper to try to prove my point.

I compared how the year-old model scored each potential deal to results from the new model trained on the latest data. There was a clear difference.

The new model predicted that custom deals with startup customers would have low profitability, but the old model thought the same deals would have high profitability.

It turned out that the old model learned two very important things:

- When a customer decision maker personally negotiates a deal, expected profitability increases.
- Custom projects are more profitable than other deals.

Both factors seemed reasonable, and the new model had the same factors. But the new model had learned something more based on recent data:

- When both conditions are true (decision maker negotiator and custom project), it's bad for profitability.

That made no sense. How could the combination of two good things be a bad thing?

I went to speak with salespeople to get some insight into what was happening. When Fortune 500 firms negotiated a custom project, the decision maker was not directly involved in contract negotiations. However, when a *startup* requests a similar project, the decision maker is almost always involved in the contract negotiations.

Fortune 500 projects are typically very profitable, but startup projects are hit or miss, often because the startup incorrectly projects future demand.

In the original data, there were so few examples of startup projects that the AI had not learned this important fact.

The old model accidentally rated startup projects high because it saw two good things—custom project and decision maker. And because at that time ManuCo did not do many deals with startups, the old AI did not see enough cases to learn the unique patterns related to startups. This, in combination with the new incentive program (which ignored the scale of

the deal), caused salespeople to close more startup projects.

Todd was right; no one had *decided* to focus on startups. The combination of the AI and the incentives accidentally pivoted the strategy of ManuCo without its leadership even being aware of it. The model was incentivizing the wrong behavior.

Jit came with me to talk to Todd about what I had found. As I explained what had happened, Todd was clearly shocked and said, "I took an AI and set up a very simple incentive program on it, and that changed my entire business strategy? How is this even possible? I didn't know this could happen."

Jit said, "Remember: AI is always at scale. It's the opposite of the way you do strategic changes today, where executives come up with a new strategy and it takes some time to take that strategy and turn it into various tactics that can be used by the frontline employees. In the case of AI, once you roll out a prediction model and an incentive program, you're directly impacting the actions of every salesperson you have. So even though you did not explicitly create a new strategy, there was an unexpected strategic implication to the incentive program you rolled out."

"So what are you telling me to do?" said Todd. "Shut it down?"

"In the near term, we'll quickly adjust the model with your help. Longer term it means understanding that AI models must be closely monitored. You can't set them and forget them."

"The very fact that people on your team combined two of Vera's models without discussing the possible ramifications with you indicates that they don't fully understand this aspect of AI," Jit explained. "And we get it; this is all new. But AI must be monitored, and specifically by people like you who know the business."

"The ability to go from strategy to execution immediately is an awesome power, and you're right to be wary of it," said Jit. "Even if that execution is perfectly on point, business conditions will change, competitors will evolve, and trends will change. Your AI will need to change too."

JIT'S TAKE

AI is always at scale, especially when it is tied to incentives or process automation.

AI isn't magic. Even the best model exists in a complex set of interactions where things go wrong in unexpected ways. It is easy to assume that AI will always solve the whole problem or tell the whole story. It won't.

We have to assume things *will* go wrong and set up checks and balances to proactively detect and fix problems as they arise. The real world is never as clean as a college lab.

Models that work today will need to be revisited. The world is changing, and AI needs tuning. You can't set it and forget it.

Myth

If AI works today, it will work tomorrow.

Reality

AI is only as good as the data it trained on. The more effective an AI, the more quickly it affects human behavior, or changes market conditions, and the more quickly it goes out of date because the data it trained on no longer reflects the new business reality.

Chapter 12

Don’t Wait for Perfect Data

In which Vera finds interesting and actionable insights from work-in-progress data, and then gets chewed out for it.

My latest project should be really fun because I am helping the marketing team for ChocoCo. They make my favorite chocolate—my only real weakness. They have free chocolates all around the office. This is the best project ever!

The best except for the fact that we aren’t making much progress yet. We have been waiting for three weeks for the data. Something seems to be wrong. Jit and I have a meeting coming up with ChocoCo’s CMO Tony Noble where we should get a full update.

Tony said ChocoCo will make an additional $100 million a month in profits once this project goes live. (I thought that might be a bit unrealistic, but Jit says it’s a reasonable return for what he’s proposing.)

Once Tony explained the scale of the data they are bringing together for this project, I started to see why Jit thought Tony’s expectations were reasonable.

"I'm getting all the signals together," he explained. "We're bringing our marketing and point of sale data together. My plan is to predict how much we will sell and optimize the promotions we run and the advertising we invest in. We've also bought data on how much our competitors are spending on advertising in each market so we can respond to their moves as well," said Tony.

I see why we're here. Without AI, we would never be able to analyze such complex data.

That is, if we ever get the data. The team aggregating all of the data is now three months behind schedule and without access to the data, there is not much we can do. Jit did convince Tony to let us start playing with the data they already have while we wait for the consolidated clean data.

The first thing I focused on is some Internet marketing data. They are running some ads on a prominent website and ChocoCo analysts said the data so far indicates that the ads are doing very well.

My AI model confirmed what the ChocoCo analysts found, but it pointed to a strange pattern. The ads are really doing well in three geographies: Texas, Louisiana, and Bangladesh. Who knew ChocoCo was so popular in Bangladesh?

The ChocoCo analysts were adamant that this was a data quality error of some sort. Click fraud is a common data quality issue where online advertisements have an artificially high click-through rate so they look more successful than they are.

I asked them whether it could be a case of click fraud but Tony had insisted they use some advanced software that analyzed the marketing data to detect bots, and the analysts had confirmed that the software had done a fantastic job of preventing click fraud.

I didn't want to run afoul of Tony by questioning his new system, but Jit insisted that we had a duty to investigate this further. If we excluded the Bangladesh data, the advertising campaign was barely profitable and if he were convinced of that, Tony would probably cancel this multi-million dollar advertising campaign.

It took some work, but we eventually figured out that both my theory and Tony's software were correct. The software had indeed prevented click fraud—but only from bots. The Bangladesh data was an example of manual click fraud where people literally sit and click on links—it is rare but it happens.

Based on their testing, the ChocoCo team had been so certain that the software had prevented click fraud that they never thought to look for other forms of click fraud that the software was not designed to detect.

I was excited about finding an insight that had potentially saved the company millions of ad dollars. Tony didn't seem very excited though.

I continued by looking at the data ChocoCo was buying on their competitors' marketing spend. The data on marketing

spend was aggregated by competitor, category, and location.

For example the data would say that a specific competitor spent $10,000 on radio ads in Detroit and $25,000 on TV ads in San Francisco that week. The data only reflected past spending because it took a couple of weeks to be collected from certain data vendors, cleaned and loaded into the system. But Tony's vision was that ChocoCo would change its promotions, advertising spend, and even product inventory at each store on a weekly or even daily basis based on analysis of this data.

Even though Tony had purchased the best data he could, upon examination the competitor data was simply too high level to be useful. For example, there are many stores in Detroit and promotions are often different for different retailers in the same city. Further, we had no details on which specific products the competitor advertised in their radio ads. Promotion and advertising decisions had to be made for each product. If the competitor advertised a chocolate bar, that might have a big impact on ChocoCo's chocolate bars but less on their mint drops.

When I trained an AI model based on the data, it seemed to confirm that the additional data we did have would add very little to the overall accuracy of the predictions. The marketing spend data was expensive to purchase and collect, so we recommended against using it and thought we had saved Tony some money. He asked to meet with Jit and me in person to discuss this further.

Tony started the meeting with, "I always believed an empty mind is the devil's playground, so I gave you some busy work while we were waiting for the clean data to be ready. But now you are confusing my team with all these 'insights' you are finding and are further delaying the project. Stop this now. This project is way too important—to the tune of $100 million a month."

I was speechless (for once), but thankfully Jit took over. Jit said, "Tony, we believe in the importance of this project and your vision. We are trying to help you deliver on your vision and start gaining the benefits sooner rather than later. Let me tell you a quick story to explain my point." Oh boy! Jit and his stories… we are getting fired today.

"We were working for a European fashion house and they wanted to figure out a discounting strategy for the clothes their stores sell," said Jit.

"Essentially if an item was not selling well, they wanted to discount it to clear out the inventory in a timely manner. When we trained the AI model, it started recommending something odd: different discount rates for different dress sizes in specific countries."

Tony looked impatient, but Jit continued.

"The client was initially frustrated because they don't discount by dress size, only by dress style," said Jit. "We determined that the AI was recommending higher discounts for larger dress sizes in countries with relatively low obesity

rates and vice versa. In countries with lower obesity rates, smaller dress sizes sold out faster while the exact opposite happened in countries with higher obesity rates. But the AI had never been provided obesity data, so how did it figure this out?"

The thought of finding results without paying for data caught Tony's interest.

"Well, although the AI did not have data on obesity, it learned the size-related sales pattern from the data and was able to use it in its predictions. The 'why' mattered only to the fashion house that wanted to understand the basis for the predictions," said Jit.

Tony interrupted, "But if they had included obesity rates in the data, that would have been more accurate, correct?"

Jit responded, "It might have. The client is now testing how much the model improves if they explicitly add that data. But notice how they are already using the predictions while figuring out ways to improve them."

"The client would have run into two problems if they had tried to come up with all the possible data sources that might potentially help with the predictions," said Jit.

"First, they would have spent months or even years waiting to get the right data together before actually starting to make the predictions that are already helping them save millions." Jit paused, letting that number sink in.

"Second, remember that the human experts never thought to put in the obesity data. The AI model trained on the incomplete data pointed them in the right direction, said Jit.

He looked at Tony. "The sooner you start analyzing the data, the sooner you can start learning ways to improve the data."

Jit's argument reminded me of something I learned in Computer Science class. There was a time when people used a waterfall approach to writing code. Experts would write detailed specifications for the product and design hundreds of UI screens before programmers started writing code.

No one codes that way anymore. Today people use agile programming or rapid prototyping where you quickly build a good enough product, get some feedback, learn, and then improve the product. These modern approaches have turned out to be faster, cheaper, and actually less risky than waterfall approaches.

"One final thought," said Jit. "In your current plan, you would essentially use your competitors' marketing spending from at least two weeks back as an input into the AI model for determining your marketing and promotions strategy for this week, right?"

"That's the plan," said Tony. "What about it?"

"Do you agree that if a market leader like you starts changing your marketing strategy, your competitors will respond by adjusting their strategy? The fact that what they did two

weeks back would affect what you do next week can become a liability. AI learns from its inputs and you are letting your competitors choose an important input to your AI," said Jit.

Tony looked like he had enough to think about, but Jit wanted to drive the point home. I could see where he was going.

"Letting others have input into an AI has backfired famously before. It's a different kind of case, Tony, but I am citing an extreme example to make a point," said Jit.

"Microsoft created a conversational AI named Tay and connected it to Twitter.[9] People started tweeting racist, misogynistic content at the AI, which it of course learned. And it learned fast: in less than 24 hours Tay was sending out completely inappropriate and racist tweets and had to be shut down. Do you want to give your competitors that level of input into your marketing strategy via your AI?"

I held my breath to see how Tony would field this ball.

"I don't quite buy that apocalyptic scenario, but I think I see your point. Show me what you can do to deliver on my vision," said Tony.

Within weeks we had deployed a "good enough" model that predicted the appropriate level of spend on promotions and marketing. Within a few more weeks, the AI was

[9] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chat-bot-racist.

recommending the best promotion to run at each store. The initial results were less than the $100 million a month in additional revenues that Tony had hoped for, but at least ChocoCo was making more money from the model than they had spent creating it.

And every month ChocoCo improves the model and the profits keep getting higher.

JIT'S TAKE

Even experienced executives fall into the trap of seeking perfection tomorrow instead of "good enough" today.

No AI model is ever perfect. There is always room for improvement as the data evolves and we better understand the problem and what the AI has learned.

Given that, why not focus on getting started fast, being open to making mistakes, learning quickly, and iterating? With a transformative technology like AI, you don't need complicated strategies for success. Because the technology is changing so fast, any plans you make will be outdated anyway.

Competitive advantage will come from advancing down the learning curve faster rather than from perfect planning. The primary job of the executive is giving the team permission to fail as long as they learn from their mistakes.

Myth

You need to bring all your data sources together and clean the data before you start training AI on it.

Reality

The sooner you start training an AI on the data, the sooner you can start learning ways to improve the data. Get started and iterate. Until you actually start training the AI, you won't know what data has value and what data really needs to be cleaned. Investigate what the AI learns and add data as needed.

Chapter 13

The Dangers of Time Traveling AI

In which Vera shares how time-traveling AI models can cause a lot of problems in the real world.

I am back at ManuCo working on a predictive maintenance project with Jeff, the Director of Maintenance. Essentially, he wants to predict which parts are most likely to fail so that they can replace the part before it fails.

Unfortunately, as part of this project, I have to supervise a team of interns who are not as detail-oriented as I would have hoped.

"It sounds like you've almost cracked the case," said Jeff. "One of the interns, Elyse, I think, tipped me off that your model is 95% accurate already."

"I am encouraged by that as well, but do note that with AI numbers that sound too good to be true, often are," I said. "That's why we always double check the models before deploying them. We will make sure we vet the model using our full process before advising you that it is ready."

"Well, let me know what you find out," said Jeff. "Any predictions will be better than following those manufacturer's maintenance schedules, which I know are conservative and costing us big time."

It's time to have a serious talk with my interns.

Elyse said, "We already cracked the case! Our model is over 95% accurate in predicting defects. Let's go present it to the client."

"I have had a few cases where I thought I hit 95% accuracy and then ended up with egg on my face. I'm curious about how you arrived at that number," I said evenly.

It turned out that they had followed a fairly standard process where they took the historical data provided to us, set aside 10% of the data for testing, trained the AI on the remainder of the data, and then tested the accuracy of the model on the data that had been set aside. So far so good. But how come their accuracy was extraordinarily high?

I looked at the variables that were the best predictors in the model and one variable called Def_Res was extremely important.

What the heck was Def_Res?

As I looked at the actual data, it became clear that Def_Res was the defect reason. If a part failed, it typically had a defect reason. If a part never failed, it did not have a defect reason.

The AI quickly learned that by looking at Def_Res, it could quickly find out whether a part had failed.

What's wrong with this picture? We are trying to predict which part would fail. The training data was naturally historical data so it contained the defect reason. But at the time when the AI model would make a prediction on new parts, the Def_Res field would always be blank because a defect reason is only specified *after* a defect has occurred. Thus, in the real world, we would never be able to use this information to predict defects.

This is a classic case of what data scientists call 'data leakage.' Think of it like insider trading using a time machine. The AI is getting to magically see the future for the data it is trained on, but it would never have access to this magic data when it had to actually predict.

I went to the interns to talk to them. "Your model isn't that accurate at all; you have fallen prey to data leakage. Did you look at your predictors to see what Def_Res was? It's the defect reason, and you don't see it until *after* a part fails."

"It's like you looked at the answer key and then took the test," I said. They looked sheepish.

"To avoid data leakage," I continued, "you have to remove every variable that would not be known at the time of making a prediction in the real world. And of course, that means you need to know what the model is identifying as predictors at the very least!"

Elyse looked a bit defensive. “So why wasn’t the model 100% accurate based on the defect reason?”

“Sometimes users never specified a defect reason even when a product failed. As such, the variable is not a perfect predictor of product failure,” I explained.

“This is a data quality issue. I know you don’t see those in your data science assignments. In the real world, data is messy,” I explained.

“And next time, please don’t talk to the customer about how things are going before I look at your work,” I said.

But I was not sure that they had fully internalized the lesson I was trying to teach them. “I’ll be back after I talk to Jeff about our results so far and reset expectations,” I said.

The next day I met with the interns again. “I’m sorry I was short with you yesterday,” I said, looking at Elyse. “Data leakage is not an easy problem. You always need to look out for it.”

“Here’s a story where it bit me in the ass,” I said. They looked interested now.

“Alice, one of my friends, was running a startup that flipped homes, buying them, fixing them up, and then selling them at a profit,” I said. “About six months after I started working for Foundation Consulting, she hired us to create an AI model for her startup.”

"The company wanted an AI to predict how much property value would go up for different types of renovations. For example, how would the property valuation change if they spent $30,000 renovating a kitchen versus adding a bathroom?"

"Flipping houses can be profitable, but it's also risky," I said. "Alice put a lot of faith in me and my model."

"The dataset I used included the characteristics of each house, the appraised valuation, any improvements made to the house, and the price at which the house eventually sold. It was about 78% accurate, and we deployed it," I said.

"Months later, Alice called me back because she could see that the business was doing worse since we deployed the AI. I rechecked my model so I could tune it and I found its accuracy on live data to be only about half of what I expected."

"I called Alice and we shut the AI down. But by that time, the AI had already cost them millions, both in the investment to deploy the model and the money it lost them after it went live." Their looks showed I had their attention (and their sympathy).

"The problem was data leakage," I said.

"Wait: You said the model was not that accurate, so it couldn't possibly have had data leakage," said Elyse.

This is exactly what I was concerned about. What they remembered from our last discussion was that the variable associated with data leakage was an exceptionally good predictor. That was the wrong lesson.

"Listen up, everybody," I said. "Did you hear what Elyse said?"

The rest of the interns looked up from their phones.

"High overall accuracy is not a necessary indicator of data leakage. Data leakage can happen with any model, anytime, to data scientists of all levels of experience. It can and will bite you in the ass," I said. "You can have fantastic predictors in your data that are not examples of data leakage," I clarified. "The reverse is also true. You can have data leakage in models that don't seem to have very high accuracy."

"The real question, the only question, is whether or not we trained the AI only on variables that would be available to it at the time of making a prediction."

I asked them point blank, "Does the definition of the variable called valuation matter in this analysis?"

"Yes," said an intern, to show me he was listening.

I looked at him. "Valuation is complicated. It is possible that the valuation represents an independent appraisal at the time the house was originally bought, before any improvements were made. It is also possible that the valuation was

calculated *after* the improvements were made to determine the price at which the property would be listed for sale. How would your analysis differ in these two cases?"

The team finally saw the point I was trying to make. "In the first case," said Elyse, "I would have this information at the time of making the prediction of which improvements we should invest in. In the second case, we would not have the valuation information at the point of deciding which improvement we should invest in. So, in the second case if we included that information in the training data, it would be a case of data leakage."

"That is right." I added, "And in this case, sometimes the valuation was done before the purchase while in other cases, especially when they bought the property at an auction, the valuation was only done afterward."

"The main reason I missed this pattern was that I had talked through a few rows of data with Alice but by chance all of those rows of data were from cases where the valuation was from before the purchase. Because I didn't understand her business, I didn't ask the right questions and thus missed the problem."

"Alice went from being a leader in using AI to having a very expensive failure on her hands that she had to explain to her investors," I said. "Her startup almost failed."

I paused to make sure they really internalized this lesson.

"So coming back to predictive maintenance, I said, "What could have helped you determine whether or not your model had data leakage?"

Elyse said, "We have to know the data better and what it means. We should have asked you for the definition of each variable to determine what it was and at which point in time each variable would have been available."

"And, to be honest, Elyse, I wouldn't have known the answer. It takes some digging to find the people who really understand what all the variables mean. That's why customer interviews are crucial when designing AI. We can't work in a vacuum. We need people who know the business and the data."

JIT'S TAKE

Data leakage is a fundamental problem in AI design, and it's harder than it looks at first.

Data leakage sounds like an esoteric technical concept, but you need to know about it because it can impact your career.

Even a well-meaning data scientist like Vera, who is trying to do due diligence to prevent data leakage by checking the data with someone else, can make this mistake. Data scientists don't know your business like you do. They need your help and your business expertise.

In other words, when your data science team tells you the model is ready, ask them to try harder to break it before

you approve it. Bring in a business expert to ask them hard questions about every single variable they're using. If they've done their job, they should be able to answer those questions.

Myth

The right data scientists and the right algorithms can create perfect AI if they have the right data.

Reality

Only a business user with a deep understanding of the underlying process can distinguish between a good predictor and a case of data leakage. Even experienced data scientists can fall prey to data leakage. Trust but verify.

Section IV

Ethical Considerations

Can a book about practical AI really discuss ethics? I believe it needs to, precisely because ethical considerations are the compass that can lead us to success in AI. AI is such a transformative technology that it has massive ability to impact our lives. In any such change, the key question becomes whether the technology will inherently empower people and positively impact lives or concentrate power in the hands of the few while the rest end up relatively worse off.

Today AI is poised to go down one of two paths. In the first path, just a few data scientists will create AI and the rest of us will have our lives changed by the predictions and recommendations generated by AI. Of course, we will try to train as many data scientists as possible, but in that world of data science haves and have-nots, most of us are unfortunately going to be in the have-not camp. The second path, the one I am a proponent of, is Intelligence Augmentation, not just AI. It recognizes that people need to be an important part of the process if we want to unlock the value of AI. An IA practitioner focuses on empowering business users by really giving them tools such that they can create, understand and update their own custom AI.

Much has been written about the big ethical dilemmas in AI—such as those related to general artificial intelligence that becomes more intelligent than humans. I am far more concerned about the everyday ethical dilemmas related to AI that we as business executives have to navigate.

In this section, first, I will focus on the question of whether AI will steal our jobs. I am sure you have seen the consulting

reports that rate jobs based on the extent to which the job could be done by an AI. But in reality, every job will be affected by AI. However, if we can guide the AI revolution down the right path, no jobs will be stolen by AI. Fundamentally AI can give people superpowers and make them more effective at whatever they do. If we focus on this aspect of AI, we can help our employees achieve more than we ever thought possible. When you create that much more value for society, there is always a way to make people better off such that they are focused on creating value through the use of AI as opposed to worried about automating themselves out of a job.

Second, I will look at how we need to consider all stakeholders when we evaluate an AI project. Any successful AI project will affect the lives of stakeholders. Moreover, because AI can be scaled almost instantaneously by pairing it with business process automation or incentives, the impact will be substantial. Let's focus first on growing the pie by maximizing the value created by AI for all of our stakeholders instead of trying to use AI as a weapon to get a larger share of the pie.

Third, I will explore the issue of privacy. Because AI often requires access to large volumes of data, people have come to regard privacy and AI as inherent tradeoffs. I reject this assumption. You can have better AI *and* better privacy. This chapter will explore the best practices for ensuring privacy while creating an AI.

Finally, I will touch upon the thorny issue of bias in AI. While most of the debate has been about accidentally introducing bias into an AI, to be honest, if we just train AI in an unbiased way on historical data, we will end up with a biased AI. This is because the AI would simply learn about the biases of our past as encoded in our historical data. Until fairly recently, there were very few Indians in leadership positions at large tech companies even though there were many Indian engineers at these companies. As a result, an AI might have easily learned the fallacious pattern that Indians make better engineers than managers. Yes, the AI is driven by unbiased math, but it is still learning from biased data. In this chapter we will explore both how to avoid bias and how to introduce positive bias to control for past biases encoded in our historical data.

Chapter 14

Will AI Take My Job?

In which Vera learns why AI needs expert business users as much as users need AI.

Jit and I have been called in by EnterpriseSoft's CEO, Maria Armstrong, for a very interesting project: helping them create an AI culture.

What do they mean by this? Maria came back from the World Economic Forum annual meeting at Davos convinced that AI would be more transformational for their company than even the Internet. She wants the perfect plan for becoming the leader in using AI in the Fortune 1000.

Her leadership team conducted an extensive analysis of various AI products and then deployed AI solutions across the company.

I asked Bill Wren, their AI project lead, how it was going.

"We managed to automate a bunch of busy work throughout the company. We've already saved $2 million from staffing reductions," he explained.

That didn't sound like creating an AI culture in my view, but I needed to know more.

"What problems are you encountering?" I said.

Bill laid it on the line. "Well, we spent way more than $2 million to get the initial savings from staff reductions."

"Second, our AI projects have stalled in my view because, some people not using the predictions being made by the AI."

EnterpriseSoft had even tried to enforce the use of the predictions by building business rules based on them. The users still found ways to work around the business rules and always had 'good reasons' for doing so.

AI was not being adopted by the company, and they needed to fix this now if they wanted to be a leader in the use of AI.

Bill gave us a simple charter, "Predict where AI will be five years from now and help us reorganize our company to maximize the benefits of AI."

Is he serious? If I could predict where AI will be five years from now, I would just go make a killing in the stock market and travel the world in style.

Jit seems very interested in this project though. Maybe it's just because he really hit it off with Bill. Apparently, both of them studied dance in college. Who knew Jit could dance?

As usual, I started by talking to a bunch of users. A common theme showed up very quickly: "The AI told me to …"

There was some clear emotion around that phrase, so I dug deeper.

I was confused that the users were saying the AI told them to do something. This AI was not telling them what to do (recommending actions); it was *predicting* what would happen. The senior leadership had assumed that the employees would interpret those predictions and take the right actions.

However, for the employees, it was easier to blindly accept the prediction and take the safest possible course of action based on it. They were running scared of AI, and not feeling empowered to balance the predictions with their business insight and domain knowledge to make the right decision.

Essentially, they did not want to risk being considered one of the people resisting the adoption of AI. While no one explicitly said it, the vibe seemed to be that following the AI's predictions might let them keep their jobs. (I also sensed that they took comfort in the fact that if a deal did not work out, they could blame the AI for what happened.)

While I was talking with this group, a tall man walked into the room and waited for a lull in the conversation to bring up his perspective.

"The AI is just plain wrong," he said. Everyone else trickled

out of the conference room, leaving me with John Stone, one of their top salespeople. "The AI simply does not know what I know about my business," argued Stone. "For example, it suggested I was giving too high a discount to one of our oldest customers."

"That AI has no idea that the customer's CEO is presenting at an upcoming conference, and we are giving them a higher discount because we really need them to say good things about our new product at this conference," he said.

"If we don't give them the discount, they simply will not buy our new product. Of course, they will still buy what they always bought from us," explained Stone. "The AI only knows that a deal will close even if I don't give the additional discount. It just has no idea that I need that deal to include our new product. If our business was a simple mathematical probability, I would not need 20 years of experience to do this job."

I presented my findings at a status meeting with Bill and Jit. Bill wasn't surprised by the skittishness of most of the business users—he'd seen it first hand—but when he heard about Stone, he said, "I am really disappointed that he said that. He is one of our most loyal employees and I have always felt that he had good judgment. But I guess when powerful new technologies are introduced, the old timers find it most difficult to adapt."

Before I could say anything, Jit jumped in, "I would not

dismiss Stone's concerns so quickly. What he is really saying is that the AI does not know all the factors that he as an expert knows."

"In one of my first consulting projects," said Jit, "I worked with an electronics equipment manufacturer who needed help with a product that had a very high defect rate."

"We used an early form of machine learning to identify patterns. We saw that the defect rate was only high in India, China, and Saudi Arabia and at five specific customers," said Jit. "If you excluded these countries and these customers, the product defect rates were not all that high."

"The product owner immediately jumped in and pointed out that those customers were in the mining industry and that the countries in question had significant air pollution problems," said Jit. "Then she said her product actually had a fan that could be easily affected by particulate matter in the air. Perhaps the fan had been the problem all along?"

"They eventually figured out that she was right and designed a workaround for environments with high air pollution. Instead of spending tens of millions on redesigning the product, the customer ended up spending thousands on a workaround for a small well-defined group of customers," said Jit.

Bill interrupted Jit, "So the AI figured out what the human could not, right? If the machine did not point out the underlying pattern to the problem, you're saying the expert would not have been able to figure it out."

Jit replied, "Exactly right. In fact, the product owner in the story made the same point herself. She had tried her best to manually figure out the pattern and had failed to figure it out because she had never zoomed into just those customers and countries."

"But, also note that the AI did not know that the five customers were mining companies, about air pollution rates in those countries, or the fact that the product had a fan which could be affected by pollution," clarified Jit. "Without these insights we would not have found the right solution. The human expert was absolutely crucial to this process."

"I get that," continued Bill, "but you could tell the AI these facts, right? I could have added the customer industry to the data for example. And, once the AI knows that, it knows it forever."

"True," Jit responded, "but the facts that you will need your AI to know will change over time. For example, we had a customer where their very basic homegrown AI was flagging Brazil as a key investment target."

"When we analyzed it, we found that the customer had indeed seen a spike in business in Brazil, but only just prior to and during the Olympics. The AI did not know that the uptick in business was related to the Olympics and was expecting the trend to continue," Jit went on. "Sure, we could make the AI aware of the Olympics by changing the data, but the key point is that some other new pattern will emerge in the future that we would have to retrain the AI on."

"Your business evolves constantly and thus the corpus of information an AI would need to know will also evolve constantly," said Jit. "You can't just train the AI once and hope it will work forever. You need a way to continuously feed in appropriate input from your experts."

Bill looked defensive. "Now you've got me worried. Are you saying AI doesn't work? I have a charter to completely automate 50% of our processes in the next decade. Are you saying that's impossible?" asked Bill.

Jit responded, "There is a huge difference between completely automating half your processes and making your people twice as effective. Both have similar theoretical financial benefits, but the latter is quite possible, while the former is probably not."

Jit continued, "AI can make your people more effective—give them superpowers, if you will—but it won't *replace* your people completely. Think Intelligence Augmentation, not Artificial Intelligence."

Bill looked a bit shaken but was clearly thinking through Jit's argument.

One of Bill's colleagues decided to challenge Jit directly. "I am surprised to hear you talk like a Luddite. Industries from restaurants to car manufacturing are completely transforming into fully automated systems. Have you seen what Elon Musk is doing with the Tesla factory automation? Why can't we do the same?"

I jumped in here, because I had just read an *Ars Technica* article[10] on exactly this point. "Actually Musk changed his mind on the benefits of extremely high levels of automation, I said. "He tweeted, 'Yes, excessive automation at Tesla was a mistake. To be precise, my mistake. Humans are underrated.'"[11]

"Tesla and previously GM actually found that while automation is useful in speeding up production, automation beyond a certain point is counterproductive," I said.

Bill stepped back in, "So Jit, Vera, you two are the experts here. How should I think about the charter I have been given to help transform our company through AI?"

Jit responded, "Well, since you are a fellow dancer, let's see if this metaphor works for you."

"If you want to lead your partner in a certain direction, you don't just shove her in that direction, right? That almost never works," said Jit.

"As a dancer, you are well aware of which way your partner is facing, which foot she is standing on, and where nearby dancers are," said Jit.

"Good dancers lead their partners in a direction that is easy and

[10] https://arstechnica.com/cars/2018/04/experts-say-tesla-has-repeated-car-industry-mistakes-from-the-1980s/. It's well-worth reading in full.
[11] https://twitter.com/elonmusk/status/984882630947753984.

safe for them to follow. They may want to lead their partner left, but sometimes it is easier to move left by first moving in a different direction that would be easier for the partner, building trust, and then guiding the partner back toward the direction you wanted to go in the first place," said Jit.

He added, "Let's look at someone like Stone. He is clearly invested in your company and has a lot of useful knowledge. But he was not part of your AI project and the process was imposed on him as a corporate mandate."

"Moreover, if the charter was described to you as fully automating 50% of your manual processes, I am sure an even more distorted version of that corporate objective would have made it out to employees like Stone. I am sure that from Stone's perspective, it feels like he is being shoved, not led."

"Employees are already afraid that AI will replace them," said Jit. "They don't realize that AI can make them much more effective and make their jobs more interesting."

"Now imagine a world where from the CEO on down, your company focuses on Intelligence Augmentation and using AI to empower each employee," said Jit. "Where you explain how AI will help you grow orders of magnitude bigger with the same size team in the next decade. Where you invite every employee to be part of this process. Where you set up a process where you solicit advice from employees on where and how your company could use AI," continued Jit.

He's really painting a picture here, I thought.

"You should give awards for the best ideas, whether or not the ideas pan out, said Jit. "You could create a center of excellence where people can bring their ideas and work with others to flesh out the ideas, train the AI, test it on the ground, and then improve the AI."

"Celebrate successes but don't penalize failures because with any new technology, some ideas won't work out," Jit continued. "With any new technology, the company that succeeds is the one that goes down the learning curve the fastest. Foster learning, innovation, and failing fast. If you can do that, if you can lead and not shove, you will be able to leverage the goodwill and domain knowledge of people like Stone to deploy AI throughout your company."

"That is the only way I know of to create an AI culture—focusing on empowering people through IA, not replacing people with AI."

JIT'S TAKE

A cynic might say that talking about making people twice as effective is just a sophist's way of saying 50% labor reduction.

The cynic would indeed be right if this was just a matter of changing how we talk. But, time after time I have found that when the executive team really changes their way of thinking to focus on how to make each employee more effective, they end up approaching the AI opportunity itself differently.

AI has such huge potential for helping our companies grow revenues, cut costs, and manage risks that it is easy to reap the benefits without unnecessary layoffs. In the real world, there is no reason to believe your workers will be worse off as you implement AI.

Soon after Henry Ford introduced automation into car manufacturing, he ended up doubling his workers' salaries and reducing their workday. Why? Because Ford workers needed to work in a different way than other workers, they had unique skills and could demand higher salaries as a result.

Moreover, imagine what happens if Stone participated in a successful AI project that eventually led to him being laid off. Would it be easier for Stone to find another job if he can talk about how he helped make his prior employer more successful through the use of AI? There are so few people with real understanding and experience with AI that other companies would be far more interested in hiring him if he was an important part of rather than a passive victim of such an AI transformation.

Of course, on the flip side, people like Stone get to choose whether they act like victims or embrace AI and use it to transform their businesses. If motivated knowledgeable employees focus on the world of possibilities that AI opens up for their companies, they can be a part of this revolution from the bottom up. Show executives how you can use AI to sell more, do more, and compete better and they won't focus as much on tactical issues like cost reduction.

Myth

People need AI.

Reality

AI needs people.

Chapter 15

Ethics of AI

In which Vera learns that almost every AI model has ethical questions that must be wrestled with.

Today Jit and I are helping Margaret Darlington, the CEO of EmpatheticInsurer, a large insurance company primarily focused on car insurance. They want to figure out which new customers are most attractive based on their probability of having a car accident next year.

EmpatheticInsurer calculates a predicted claim payout number for every member it insures. If a new member's expected claim payout is lower than her insurance premium, the insurer makes some money from the member that contributes to the insurer's operating costs and profits.

The problem seemed simple. If we can better predict expected claim payout rates, we can better price the policy and thus help EmpatheticInsurer make more money. My team and I quickly created a fairly accurate way to predict the expected payout for a new member and went to review the model with Jit and Margaret.

Typically in such a model review, Jit suggests ways to maximize the benefits of the model based on the customer's cost-benefit

characteristics and operational constraints. But this time he was uncharacteristically quiet, staring out the window.

Eventually he looked up and said, "Margaret, have you thought through what your organization is going to do with this model?"

Jit said, "Ethical questions are embedded in the way we create any AI model, and I am sure you want to surface those questions up front. For example, if we focus on expected claim payout, we are in essence focusing on excluding unprofitable new customers. Such an AI would do nothing to help your current members. Should our focus not be on improving the results for your existing members as well as finding a truly differentiated way to create value for new members?"

Margaret said, "It sounds like you are talking about a more strategic approach towards AI than the tactical problem we wanted to start with. I am a bit confused about the ethical question here."

"It's all in the way you frame the business problem," he said. "What exactly is your business goal?"

"We want to predict and reduce claim payout costs," she said.

"But is that really your objective? The easiest way to reduce claim payout would be to not write any new policies," said Jit. "Given your brand, I know that serving your members is the top priority."

"If we just predict next year's costs for a specific potential client, what is the easiest way for the company to use this information? Couldn't the financial incentive cause you to avoid accepting new members with high predicted claim payouts?" said Jit.

"But what's the alternative?" said Margaret. "We are in a very competitive market and if we don't figure out a better way to price policies, then we will quickly lose out to our competitors who are already rolling out AI to price policies."

That's exactly what I was thinking. I'd seen the numbers and the trends.

"Really what I'm talking about is adjusting how you frame the question. Let's imagine a model where we recommend actions that would significantly reduce predicted payout costs next year for your existing members," continued Jit.

"For example, for a certain member, the system might say that if we can get this person to use a device that warns them if they go over the speed limit, then their expected claim payout would be significantly lowered next year."

"While most members would benefit from improving their speeding habits, not all of them would benefit to the same extent. My wife for example has always been careful about speeding and originally such a device would not have been very useful for her. But as soon as she got her new electric car, due to the absence of the engine sounds, she found herself accidentally going over the speed limit. We needed to turn

on the speeding warning on the car so that she could avoid speeding without realizing it. By prioritizing members with the highest expected cost reduction and giving them focused help such as free software on their phone that warns about speeding or encouraging them to buy cars with such features, EmpatheticInsurer can create a win-win. Essentially you would be helping your members avoid car accidents while systematically reducing your costs to serve them."

Margaret said, "I like your approach. A big part of our mission here is to improve member experiences so something like this would fit well with our mission. But changing people's behavior is not easy. It is easier to simply take on less risky members than hope to change the behavior of our members to reduce costs. If the members don't change their behavior, we could lose a lot of money."

"This is where we come back to your mission and the way we design the model," Jit said. "I believe that if we think through the impact on every possible stakeholder, then the ethically correct decision is almost always also the most profitable decision."

Jit continued, "If you just focus on predicting expected payout rates, and then set prices accordingly, this will definitely affect your new members. Some customers may see their insurance costs increase because of new factors that you are considering with the AI. But, this is in a way an adversarial relationship with your customers. You are trying to maximize economic benefits for yourselves but every

extra dollar you make is a dollar your customer spends. You would not have fundamentally created new value for your members through the use of AI, and you would be in the world of a price war powered by AI. As you change your pricing based on AI, your competitors would change their prices in response or in parallel due to their independent adoption of AI. Such price wars are not new. Every time people have access to better prediction techniques, they end up trying to improve their pricing."

"But AI is fundamentally different. It can actually deliver personalized recommendations to each of your members to help them reduce their risk of accidents and thus your expected payouts. In essence, the AI driven recommendations would help you improve the insurance risk of the customer from what it was at the time when you competed for their policy. This improvement is not visible to your competitors at least until the customer applies for a new policy. Why just focus on a price war that would be visible to your competitors instead of privately creating real benefits for your members and reaping the benefits?"

"Think of how you can use AI to create the greatest overall value first. When you design your AI to recommend specific interventions to reduce the risk of accidents for members, you make it easy for users to leverage the insights to save money by doing something that benefits everyone."

"When you only predict next year's claim payout, the easiest way to use that information is to avoid serving the members

who are most likely to cost more next year, thus invalidating some of the core value proposition of insurance."

"Thank you for helping me begin to think this through. I would like to go back to some people on my team and talk about this more before we proceed. There may also be synergies we can explore. We have some accident avoidance initiatives going on that might be able to work in tandem with the new AI that would place the AI more firmly in the win-win area," said Margaret.

Margaret seemed satisfied, but I had questions of my own I needed to think about.

I had always believed data science and AI are just algorithms. Algorithms are not good or bad; how people use them is good or bad. However, here the nature of the algorithm I created itself makes one kind of action more likely than another. Could I then claim the ethical purity I had always maintained?

I now felt burdened by the responsibility to raise ethical questions with clients. I wasn't sure that was a fair responsibility for the data scientist. I said, "Jit, I had completely missed the points you raised. I would have been happy to deliver an accurate pricing model and declared victory. What should my role have been in the ethical process you discussed?"

Jit responded, "As we approach the design of AI, we need to think about how we can make the 'good' use of your model

the easier use of your model."

"Driving with your seatbelt on is a good thing. That is why we require seatbelts in cars and have alarms go off if we don't wear our seatbelts. Imagine if people had to pay extra for seatbelts or if there was no alarm to remind us about putting them on," Jit continued.

"A pricing model was a perfectly reasonable thing to deliver. But it could have led to suboptimal decisioning by the customer. Thus, we helped the customer see the bigger picture and proposed a different way of using AI that created more benefits for the customer as well as society at large," he said.

"But this is not only on your shoulders. You should help the customer understand the possibilities of AI because you are the technical expert there. But the final responsibility is always on the shoulders of the business user. AI is a fundamental shift that can create huge new sources of benefits. It is not a zero sum game. The ethical business executive will focus on growing the pie first as much as possible and then take a good slice of the bigger pie. The tactical executive will focus on their percentage of the pie instead of on growing the pie. The beauty of the value-creating power of AI is that the exec focusing on growing the pie almost always does better than the one only focusing on their own slice of the pie."

JIT'S TAKE

There are inherent ethical questions in every AI project. You may not see them at first. Ethical questions don't come with

blaring alarms. They creep up on you by inches.

Ethical questions must be surfaced and explored up front. If we don't think those questions through and make the right choices, we can inadvertently create AIs with negative ethical consequences. Even if we create such AIs inadvertently, we can't run away from our responsibility.

In the EmpatheticInsurer case, the results of the AI can literally determine who gets access to affordable car insurance and who doesn't. Sometimes the eventual ethical problem turns out to be something that we could not have anticipated. But at the very least, we need to partner closely with people who know the business so that together we can think through the ethical implications of the way we designed and implemented the AI.

Keep in mind that the right thing to do—the thing that is right for your brand and your company—is almost always the most profitable thing to do. If you can't see it, I invite you to take a broader look at profitability.

Myth

Ethics in AI is mainly about the science fiction future where intelligent robots take over the world.

Reality

Every AI project has an ethical component. You are always affecting the lives of stakeholders and you need to consider how AI will affect each of them.

Chapter 16

AI Protects Privacy

In which Vera sees that privacy and AI are not a tradeoff.

Jit and I are at a leading hospital system, MediChain, to design an AI that protects patient privacy.

MediChain's Chief Privacy Officer, Alex Melnick, brought us in to look at their current predictive analytics practices and design an AI-based approach that ensures the privacy of patient data so he can sleep better at night.

He explained, "MediChain is trying to predict patient length of stay and readmission rates. We don't get paid more if a patient has a longer stay, so we have an incentive to treat patients quickly and free up beds for others. However, we are paid less if the patient is readmitted for the same disease."

This wasn't news to me. Reducing readmissions is an important metric in healthcare.

I asked for more information. "Alex, can you explain how you're looking at readmission patterns now and why you are worried about patient privacy as a result?"

He explained, "Today MediChain analysts search patient

data to find patterns. By its very nature, searching requires analysts to have access to patient data and we are always concerned that analysts might inadvertently compromise patient privacy."

I added, "Could you please explain exactly what the analysts are doing when they search patient data?"

Alex nodded and started in with his explanation. "Let's say our analysts are looking at the medical outcome for Bob, an older white man with chest pains, who had a cardiac arrest and stayed at a hospital for three days while he was treated. How do we figure out whether three days was too long or too short?"

"Well, we could search our database for all patients matching Bob's characteristics and compare his length of stay to that of similar patients, said Alex. "But we could not get too specific—for example if we specified both the doctor and the nurse who treated him, there might be too few people matching the search criteria and there would not be enough examples to compare his case to. Our search has to be specific to the patient, but not too specific."

"Let's say we included the hospital name in the search, and it turns out people like Bob stayed at that hospital between three and four days. Based on that, we might think Bob's outcome is perfectly normal. However, what if similar patients in other hospitals typically stayed just one day? If my search is too specific—because I include the hospital—I may not

notice patterns that highlight problems for MediChain."

As Alex explained what the analysts were doing, the privacy problem became clear. Essentially if they were searching the database of patient outcomes, they could see confidential data of the patients. But I was still missing something here. "Alex, can the analyst see personal data like patient names and phone numbers for the people they are comparing Bob to?"

"Of course not," Alex responded, "but that is still insufficient to preserve privacy. For example, there may be only one 76-year-old male patient in a specific hospital. Thus, if I want to get access to the private patients records of such a patient, I just have to search by those characteristics and if I only see one matching record, I have just breached the patient's privacy."

Alex concluded with a specific challenge, "I have read a lot about the tradeoffs between privacy and AI. But I can't trade away privacy. So here is the key question: Can AI help me deliver better analysis while actually improving privacy?"

Jit and I live for such challenges.

After a battle to get all their patient data, I implemented an AI that looked at all of the data to predict expected length of stay for any specific type of patient. So, if we wanted to check whether 'Bob' had stayed longer at the hospital than expected, we would just predict Bob's expected length of stay and if that was significantly different than the actual

stay, then we knew there was a potential problem.

I now had to make a case for why this AI was better for privacy than the manual search-based approach.

I explained to Alex, that once the AI is trained, it doesn't need access to the data it was trained on. The AI actually learns probabilistic 'drivers' based on the raw data and predicts using that information. The AI learns about categories of patients (like older white, male patients admitted after a heart attack) in a very different way. It doesn't just search for patients with these characteristics. It learns about length of stay patterns for old people, white people, people with heart disease, then old white people, old people with heart disease and so on.

When we ask it to compare whether a specific patient's health outcome was as expected, the AI crafts a hypothetical patient just like that person based on everything it knows about every patient and then compares outcomes.

Depending on the type of AI being used, this information may be stored in various ways—for example as a table of weights and correlations—but the raw data that the AI was trained on is not required anymore.

Alex understood my point. "So even my analysts would not have access to the patient level raw data anymore? This is great. But, remember the thing I said about search patterns getting really specific and accidentally identifying private data? Wouldn't the AI suffer from the same problem such that

if there was only one 76 year old Native American patient with heart failure in my hospital system, it would store the health outcome for that patient among the information it stores to enable the predictions?"

I clarified, "The information that is being stored to enable specific predictions is information on large groups of people, and we can easily restrict the AI so that it does not look at very small groups of people. Essentially if a combination of variables gets so specific that there are fewer than say 10 matching records, we can tell the AI to not store information on such a small group."

"In fact, even if hackers got access to all the information the AI stored, they would not be able to find out any information about a specific individual—such as whether that individual has cancer. The hacker would be able to see the probability that people like that person have cancer, but that probability is driven by what the AI learned about large groups of people. This does not breach an individual's privacy. All of the benefits of AI without the privacy risk of existing manual processes—delivered."

Before I could take an actual or metaphorical bow, Alex interrupted with, "But now I have to copy the data to a new location for the AI to train on, right? That is still a potential privacy problem."

Jit came in with an assist there, "With the latest techniques, AI can be trained where your data is, even if it is on your private on-premise servers. You may need to add some

temporary servers to train the AI on the data, but once the AI is trained, it can be deployed in a completely different environment. But things get even easier if the data is already securely on the Cloud. In that case, the AI can be trained securely within the same environment without ever making a second copy of the data."[12]

JIT'S TAKE

A lot has been written about the privacy implications of AI. Whenever a new technology is developed, there is a lot of fear about it.

Regulatory impact is often a fertile area for fear mongering because executives are especially concerned about being fined or going to jail as a result of violating regulations. When cloud computing was in its infancy, there was similar fear mongering about the privacy and security implications of the cloud. Today, most CIOs would admit that the computing infrastructure at cloud companies like Salesforce, Amazon and Google is almost certainly more secure than their private on-premise servers.

Why is that? Because a cloud company's entire business

[12] For example, if your data is in Amazon Web Services (AWS) for example, we could use Amazon Sagemaker to train an AI on that data within the same secure AWS environment. The data would never pass outside the secure AWS environment. Once the AI is trained, its specifications can be taken out and deployed elsewhere in AWS or any other environment without risking the security of the data it was trained on. Other cloud platforms have very similar capabilities.

depends on maintaining customers' trust in their data security. Thus, they invest in security as a core competency.

For any company offering AI software or services, privacy and data security needs to be a core competency. Modern approaches to training AI on secure data take an almost Data Clean Room approach where the only thing that comes out of the Data Clean Room is the AI. The raw data never comes out of the Data Clean Room. Make sure your AI system is designed from the ground up with privacy in mind.

Myth

To get AI, you need to sacrifice some privacy.

Reality

AI actually reduces privacy risks.

Chapter 17

Unbiased AI

In which Vera learns that AI is not always the right solution.

Today Jit and I are working with EthCredit, a midsized credit union that built its reputation on its ethical behavior. Given what I know of their business practices, I was genuinely surprised by the problem they posed to us. They asked us to determine whether their AI is discriminating based on race and gender when it recommends whether EthCredit should approve specific loans.

Norm Robbins, their Chief Ethics Officer, explained, "We have always taken pride in the way we invest in our community and in not discriminating against disadvantaged borrowers."

"In fact, we have a track record of better loan repayment rates from some of the more disadvantaged borrowers. We regularly train our employees to prevent unconscious bias from creeping into our decision-making," said Norm.

"Recently we initiated a project to completely eliminate human bias—letting an AI independently evaluate loans without any knowledge of the applicant's race, gender, and similar factors," said Norm.

Jit and I exchanged a look because we knew that AI can infer excluded data.

"We thought AI would be less biased, but as far as we can tell, the AI is rejecting loans from women and minorities that our loan officers would have typically approved. If a human employee had taken the same decisions, we would have reviewed their work to determine whether there was evidence of race or gender bias," said Norm.

"Then I read an article[13] in *ProPublica* that uncovered significant racial bias problems with an AI-based risk assessment tool used by courts for parole decisions. I could see that it's more complicated that I had thought," Norm said.

"And to make matters worse, some of our executives are questioning whether the AI is actually correct and saying that we took our anti-discrimination training so far that our employees were biased toward disadvantaged applicants to such an extent that the fundamentals of our banking business are at risk," he said.

"This experience is making us question the way we have run our business for decades. We need to figure out whether the AI is right or not."

[13] See http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. This is an important enough topic that I would strongly suggest also reading. http://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing and http://motherboard.vice.com/en_us/article/4x44dp/ai-could-resurrect-a-racist-housing-policy.

Jit responded, “I read that ProPublica article too. It seemed to show clear racial bias on the part of an AI that was being used to predict which criminal defendants were most likely to reoffend.”

“However, because the details of the AI itself are proprietary in that case, it is impossible to definitively assess what is going on,” Jit lamented.

“Speaking broadly, an AI may be biased against a subgroup if it was not trained on sufficient examples from that subgroup. A famous example involves a case where an AI was trained to detect pictures of people as opposed to animals based on a dataset that mainly had pictures of white people. The AI thus had not learned that people can be darker skinned as well. When presented with a picture of an African American male, the AI predicted that he was an animal.”

Norm looked shocked. I wished this story wasn’t true but it is actually a well-known cautionary tale in this space.

“Such stories horrify us,” Jit continued, “but the key point here is that the bias is in the data not the AI. If your AI is trained on sufficient unbiased data on each subgroup, it should be fairly unbiased.”

“But when you say that your AI was not aware of the race or gender of your applicants, do you mean that you simply removed these variables from the data that was presented to the AI?” I asked.

"Yes," said Norm. "I asked specifically that those variables be removed. If the AI is simply not aware of gender, for example, then it can't be biased by it, right?"

I looked at Jit. I knew he would not answer that question until we had dug into the details a bit, but I could already see a gaping hole in Norm's reasoning. BigBank had also run into this problem, and they were skittish about AI predictions as a result.

Removing a variable like gender from the data does not remove gender bias because other variables may indicate the gender of the applicant.

For example, schoolteachers are disproportionately likely to be female. Thus, the profession of the applicant can imperfectly indicate gender.

Certain shopping websites impute an anonymous visitor's gender and age based on which page the visitor came from, the search terms they used, what pages they clicked on, and even what device they used to visit the page.

Such approaches don't perfectly indicate the visitor's gender. But they are good enough that if the goal is to eliminate gender or racial bias, getting rid of those variables is not a sufficient solution.[14]

[14] Data scientists have advanced ways of addressing this problem. One approach initially leaves these restricted variables in the data, trains a model, and them avoids using these restricted variables in the actual predictions. One way to think of this is that because the gender variable

After talking with Norm about his concerns, we proceeded to talk with as many of his colleagues as we could.

Ralph, an experienced loan officer said, "I can't see how an AI can do what we do. For example, we don't want to discriminate based on gender. So, if a female applicant has a gap in her employment history, we ignore it if it was maternity related. This is just one example of hundreds of rules of thumb we use to avoid discriminating. How would you teach the AI such patterns?"

Another loan officer, Janine, expressed her concerns. "If AI is unbiased, and learns just from the data, then the higher-ups have no idea the can of worms they are opening up," she said. "For example, race may be tied to incarceration rates and worse health outcomes. These directly affect the financial stability of our applicants."

"And many if not most of the bankruptcies I see in our community relate to unexpected healthcare costs," she said. "If I was truly unbiased, and decided just based on the facts, I might easily take very reasonable decisions that would be considered racist or at the very least favor wealthier clients. It is

is in the data, the AI ascribes most of the gender-related signal to that variable. This means the gender-signal that was in the employment data is now mostly ascribed to the gender variable and not to the employment variable. Then, if we just don't use gender in the prediction, we are still using the predictive value of the employment variable but without the associated gender effect. Sounds complicated? In reality, this is even more difficult than I described and there is a lot of art as opposed to science in this process.

precisely because I know the difficulties faced by the members of our community, and how a small loan can turn around the lives of some of our members, that I am biased toward lending money to people we would consider disadvantaged."

"This is not a head problem; it is a heart problem," she insisted. "I am really worried that this AI nonsense will highlight to the executives that some of our decisions can't be justified by a machine. The thing is, we are profitable. It might not add up, but it works."

"If the executives forget about the business benefits we get from our reputation of doing good, then we might destroy what made our bank special in the first place," she said.

I walked away from these meetings feeling disheartened. The concerns raised by these loan officers were valid. While I could see how I could potentially train an AI to recognize certain patterns—such as the maternity leave example—it would not be easy to train it on all of the kinds of rules of thumb the interviews highlighted.

I went to Jit feeling defeated and said, "I already have a list of over 100 rules of thumb that these loan officers are using to ensure the bank lives up to its commitments to the community. This is not even an exhaustive list."

"Many of these rules are contrary to things an AI would learn if I trained it on the data. I can try adjusting the data to reflect some of these rules, but I simply can't see how I can avoid unexpected interactions between the adjustments

I would have to make," I said. "This would be an extremely fragile AI and it would require months of painstaking manual programming to prepare the data to train the AI."

Jit asked me a strange question, "Why are you assuming you need an AI to address what the customer has asked for?"

What a question from Jit of all people! I am an AI expert. I create AI models for a living. The customer specifically called us in to give them an unbiased AI solution. Were we in the same meeting or even on the same planet?

Of course, I didn't actually say that. After all, Jit has a say in my upcoming evaluation for a promotion. The shocked look on my face said enough.

Jit said, "Let me handle the customer. Could you please look into a few things for me? When you say there are more than 100 rules of thumb used by loan officers, are these rules written down somewhere? How consistently are they applied? Does Norm know about—and agree with—all of these rules of thumb?"

Off I went to conduct another set of interviews. It turned out the rules were not completely formalized. While there were several consistent themes and the goals were usually consistent, each loan officer had a slightly different approach to how they applied their 'judgment.'

This was in direct contrast to how well-defined the process of generating a loan score was. But the score was an input into

the final loan decision and the officer had a small amount of leeway in approving loans that matched the ethical mission of the bank.

Unfortunately, this small leeway still had a significant impact on which loans would get approved and my analysis showed that there was too much inconsistency in how officers applied their judgment.

We managed to rationalize the rules to a set of 40 that most of the loan officers agreed upon. When we tested these rationalized rules with Norm and his team, they approved almost all of them.

Finally we had a list of 36 rules that everyone could agree on. Our next step was more challenging: we had to explain to the client that they should not use AI to solve this problem. I am just glad it is Jit who has to handle that discussion.

Jit started the meeting with the EthCredit leadership team by walking them through some of the basics, including the fact that just removing variables like gender from the AI training data does not remove gender bias.

Then he waded into more dangerous territory: “We were asked to help build an AI that would be unbiased. However, it has become clear to us that EthCredit actually needs a consistently biased AI.”

“Essentially your problem is that you want your loan approval process to bias toward certain types of applicants where a

cold-hearted unbiased evaluation of the loan application might have meant the loan would be rejected," he explained.

"This is because you understand that the success of your business depends on more than just loan repayment rates. Each of your loan officers is introducing a positive bias into the approval process. They are just introducing the bias in an inconsistent manner," he said.

"Now, we could try to train the AI to learn their biases from the historic loan approval data, or adjust the data or algorithms to build bias into the AI, but you would have no clear understanding of or visibility into the bias introduced into the AI. Why not just have a set of 36 pre-approved rules that simply adjust the final loan approval score in a consistent way and solve the problem that way?"

As I expected, the client was not very enthusiastic about Jit's point. Norm said, "Are you saying there is no way to create an AI to solve our problem, or are you saying you are not capable of creating an AI to address this problem? Should we just ask a different firm to help us?"

Jit responded, "Of course there is a way to train an AI on this. If you just train the AI on your historic loan decisions, it will pick up some of the rules of thumb your loan officers used."

"But remember: the officers themselves were not consistent. Moreover the adjustments were a small part of the overall approval decision, most of which was determined by the loan score which is based on simple math and not based on

your ethical goals," he continued.

"If you train an AI on this data, it may ignore the positive bias you are trying to maintain, both because the bias was inconsistently applied historically and because it was a much smaller part of the overall decision," said Jit.

"If you really want to, you could create an artificial dataset where you take historical loans and take theoretical approval decisions based on these 36 rules you have agreed upon and then train the AI on that," said Jit.

"The important thing to note is that you will be using a complex way to solve a simple problem. Moreover, if you ever have a regulatory review, it is far easier for regulators to review and approve a simple set of 36 rules than understand the inner workings of a biased AI."

The regulatory argument was the clincher, but I could see from the body language of the executive team that they were beginning to appreciate Jit's broader point.

"So you are saying we don't need to use an AI when a simple set of rules will suffice?" asked Norm.

Jit replied, "I couldn't have said it better. AI has its place but it doesn't need to be used everywhere. Sometimes simple solutions are good enough."

JIT'S TAKE

Human judgment is used everywhere in business. For example, we may invest in new markets or products even if they are not initially profitable.

If we trained an AI on our data without thinking when and how we apply human judgment, we could end up with some dangerous outcomes.

For example, if we have lower expected probability of winning deals involving a new product, the AI may recommend that we give up on sales efforts for that product and prioritize pursuits involving our more established products where we have a higher probability of success.

This could cause us not to invest in new markets and products and lead us down the path of being disrupted. Business is always more complex than a simple number that can be predicted by an AI. We need to figure out how to explicitly understand and address these moments of human judgment as we design AI systems.

We also have to acknowledge what AI can't account for. Values, brand loyalty, goodwill—anything that is important but difficult to quantify goes beyond the purview of AI.

The more human experience is required for a decision, the less likely it is a good candidate for AI.

Myth

AI reduces human bias because it is not human.

Reality

AI learns from the data so it learns human bias. If you just want to reduce bias, AI may not be the right solution.

Conclusion

It is not lost on me that in the last chapter of this book on AI, Jit recommended a solution that did not involve directly using AI. People often claim AI can be used anywhere and can improve anything. Once you have a hammer in your hand and go looking for nails, everything will look like a nail. It is very important to build your own intuitive grasp of where and how AI can create value for you. If AI is your answer to every problem, look in your toolkit for more tools than just that one hammer.

AI IS ALWAYS AT SCALE

This is especially true of AI because AI is inherently different from every other technology we have seen before in terms of how quickly it can affect our businesses.

Everything we have done to date has had a human element to it. The PC, enterprise software, the Internet: all these technologies required a lot of training of humans, a lot of change management, and years went by before the technology was widely adopted. Because of the natural friction introduced by the human element, if your idea were really bad, it would probably get found out before it was adopted widely.

The interesting thing with AI and the combination of AI deployed into enterprise applications and business

processes is that we could literally turn on an AI system today and instantaneously affect how our users work. For example, you could simply stop showing your salespeople sales leads that an AI believes are unlikely to close. And you have instantaneously, in a span of an hour, fundamentally changed the way your business is run. And that is what is scary about it.

Remember: today AI only learns from what it has seen. There are certain techniques where the AI might try a good educated guess, but it still fundamentally will not know how to react to something it has never seen before. So when you turn on AI at scale like that, when you fundamentally change your business overnight and you're actually hiding information from users, it can very quickly have an unintended negative impact on your business.

Note that at a very fundamental level even the data scientists who designed the AI don't completely understand what it is predicting based on. My favorite line in the Vedas comes at the end of the Hymn of Creation,[15] a long poem about how the world was created. It ends by questioning whether the philosopher poet actually knows how the world was created. It then goes on to ask whether even 'the one above' (God) knows exactly how the world was created. Keep that sense of skepticism and humility in your heart as you evaluate and use AI.

[15] 129th hymn of the 10th Mandala of the Rigveda

IA VS. AI

Today the field of AI can go down two divergent paths—magic AI where the AI tells us what to do without much explanation—or human empowerment where the focus is on human-machine synergy. Think of it as Artificial Intelligence (AI) vs. Intelligence Augmentation (IA). If we go down the path of magic AI, we will have fundamentally reduced our ability to transform businesses at scale by leveraging the best AI in parallel with the best intuition and domain knowledge of our people.

In this book, I almost always recommend IA first. There are good reasons for keeping a human in the loop and treating AI as something that has to be continuously monitored and improved.

When we truly understand that all AI will fail, and we design our systems for resilience and learning, we start setting the groundwork for the real potential of AI. The true story of AI in business is not about how we can successfully deploy individual models at scale and benefit from each model. The true question is how we will reimagine the way we do business in light of the superpowers AI can give to every employee.

Business today is organized by the constraints of yesterday. We need to rethink those constraints. Some of my favorite projects spanned corporate silos such as sales, marketing, logistics, and finance. Humans can only handle a certain level of complexity. As such, once a process starts spanning different business functions, we quickly approach the

limits of our business knowledge and our ability to handle the complexity of what is being analyzed. IA can however help users from different silos collaborate across silos to reimagine the way we do business. That is true human-machine synergy as a catalyst for societal change.

That is human empowerment through IA.

A Practitioner's Checklist for AI

Whenever you undertake an AI project, you should think through the following questions:

- □ **Objectives:** What exactly do you want to achieve here? What is the business outcome you want to impact? What are the other business outcomes you need to consider? (For example, if your business outcome is maximizing win rates, you may also want to ensure that people don't discount so deeply that the deals become unprofitable.)

- □ **Staffing:** Do you have at least one person for whom deploying AI is their primary job? It is perfectly fine if most people working on AI projects have other business roles that are their primary jobs. However, someone needs to be responsible, empowered, and incentivized to systematically deploy and promote the use of AI.

- □ **Actionability:** Can you actually affect your business outcomes? Are you clear about which kinds of

actions you can take to effect change and are the corresponding actionable variables included in your model? [Skip if your AI problem does not relate to changing a business outcome.]

- □ **Stakeholders:** Have you thought through everyone who will be affected by your AI? Are the results of the AI beneficial for all stakeholders? If not, under what circumstances are the outcomes likely to be negative and have you thought through your worst-case scenario there? Have you involved every kind of stakeholder in your AI project? If not, have you thought through why you excluded some types of stakeholders and considered whether those types of stakeholder might have information that could affect the effectiveness of your AI?

- □ **Ethics:** Have you designed the AI to make the right thing to do the easy thing to do? Are you trying to genuinely empower your stakeholders by using AI? Have you addressed any potential negative impact on stakeholders? What level of negative outcomes are you willing to accept for each type of stakeholder? Have you set up a way to monitor such negative outcomes and created a plan for addressing the negative outcomes if the acceptable thresholds of harm are exceeded? (For example, if we are trying to reduce discounting, there is a chance that we will lose some deals and our salespeople will lose commissions on such sales. What level of reduction

in sales is acceptable as we focus on increasing profitability? How do we plan to compensate our salespeople for their loss of commissions?)

- □ **Data:** Have you collected as much data as you can reasonably collect? Have you thought through whether the data includes your most common scenarios? Are there any orphan scenarios? (For example, you forgot to include data from a specific country.) Have you laid out the data at the right granularity for what you are trying to achieve? (For example, if you want to predict Customer Lifetime Value, then each row of your data should be a customer, while if you want to predict Annual Spend, then each row of your data should be a Customer in a specific Year.) Have you excluded data from transactions that do not represent normal behavior? (For example, did you exclude data from the time there was an Olympic event in the country?)

- □ **Models:** Do you need one model or many? What type of model should you use? Do you need real AI, or is a rules-based system or a simple trend analysis good enough for you?

- □ **Accuracy:** Is the model accurate enough for the granularity of action you can take? How would you benefit from making the model more accurate? Is the model so accurate that you should be worried about overfitting the model to the past behavior to

such an extent that it does a worse job of predicting the future? Did you think through patterns like data leakage that can make a model look more accurate than it is?

- □ **Bias:** Have you minimized human bias? Or if bias is required for your use case, have you clearly articulated the kinds of bias or rules of thumb you have incorporated into the process?

- □ **Explanation:** Did you do your best to understand what the AI has learned? Have you reviewed how different predictors may interact with each other? Can you explain what the model is doing at a high level to any executive?

- □ **Feedback:** Have you set up a way for end users to provide feedback on the AI? Is there a systematic way to review and act upon this feedback?

- □ **Testing:** Did you treat this process like a penetration test? Did you deliberately try to break the AI by testing it on real transactions that the most cynical human experts believe it might do a bad job predicting?

- □ **Monitoring:** Are you monitoring the model to make sure it remains accurate? Are you conducting ongoing A/B testing by not using the model on a subset of the data? Are you running multiple models (champion-challenger) to continuously look for opportunities to improve the models? Will you notice if the model

goes out of tune for only a subset of the data (for example, for a new product that was just introduced)? Do you have a way to retrain the model if it goes out of tune?

- □ **ROI:** Did you clearly define how you will calculate the ROI for this AI? Did you set up a process for calculating and independently auditing this ROI? Where possible, are you conducting A/B testing to scientifically prove the financial benefits created?

- □ **Learning:** Are you creating a safe space for people to try different approaches with AI? Are you failing fast and learning from those failures? Have you set up informal and formal ways for AI practitioners to communicate their experiences and lessons to others?

- □ **Evolution:** Are you rethinking the way you do business in light of what you are learning from your AI initiatives?

Join Us on Our Mission

Thank you for reading this book. Please visit www.aible.com/ai-book to share your feedback and contribute *your* take on any chapter.

About the Author

Arijit Sengupta is the Founder and CEO at Aible and the former Founder and CEO of BeyondCore (a Salesforce company). Arijit has guest lectured at Stanford and other universities; spoken at conferences in a dozen countries; and has been written about in *The World Is Flat 3.0*, *New York Times*, *San Jose Mercury News*, *Harvard Business Review*, *The Economist*, and other leading publications. Arijit held leadership positions at several big data, cloud computing and e-business industry associations and previously worked at Salesforce, Oracle, Microsoft, and Yankee Group. He has been granted seventeen patents. Arijit holds an MBA with Distinction from the Harvard Business School and Bachelor degrees with Distinction in Computer Science and Economics from Stanford University.