UNIVERSITY of VIRGINIA | DARDEN Business Publishing

UV8553
Giving Voice to Values
Rev. Dec. 6, 2022

Move Fast, but without Bias: Ethical AI Development in a Start-up Culture (A)

"Promoted?" repeated Taylor, incredulous at the words that Darnell, Catalise's chief product officer, had just said.[1]

"Absolutely, Taylor," Darnell responded. "You've really proven yourself as a product manager here at Catalise, and we know you'll do great work as a senior product manager." Darnell ended the call by briefing Taylor on her salary adjustment and setting up a time to connect on the first project in her new role, which Darnell hinted would be a big one.

Taylor closed the video chat, spun around in her chair, and let the news sink in. As a 29-year-old woman, Taylor would be the youngest senior product manager at Catalise, and the only woman. Eighteen months before, when she had joined the start-up, Catalise was still in its infancy, selling investors and customers alike on the far-fetched idea that artificial intelligence (AI) software could diagnose patients' mental health conditions by analyzing their speech patterns. Since then, Catalise had launched Catalisence, a voice-based diagnostic tool for schizophrenia that had been adopted by an estimated 40% of mental health clinics in the United States. Catalisence's success propelled over $300 million in venture capital investment for the company, attracted partnership opportunities from prestigious hospitals and universities, and fueled rumors of a Catalise IPO in the near future. Company leaders were focused on bringing more diagnostic products to market in order to keep the momentum going.

The week after her promotion, Darnell briefed Taylor on her first project as a senior product manager. Catalise was gearing up to launch Catalisten, an AI-based software that could diagnose individuals with major depressive disorder (MDD) by analyzing their speech. Because of the prevalence of MDD—approximately 21 million Americans experienced a major depressive episode each year[2]—Catalise expected Catalisten to be its most profitable product for the next several years. As a result, leadership sequestered Catalisten developers from other product teams and research partners to avoid any competitive leaks. Catalisten was 80% market ready, Darrell estimated, but the team needed Taylor to coordinate the final steps before launch. Darnell also

[1] This material is part of the *Giving Voice to Values* (GVV) curriculum. The Yale School of Management was the founding partner, along with the Aspen Institute, which also served as the incubator for GVV. From 2009 to 2015, GVV was hosted and supported by Babson College. This case study is made possible with a grant to the University of Virginia from the Deloitte Foundation, which supports education through a variety of initiatives that help develop the talent of the future and promote excellence in teaching, research, and curriculum innovation. The case study may contain references to the content of third parties ("Third Party Content"). Third Party Content is not monitored, reviewed, or updated, nor is any Third Party Content endorsed by the Deloitte Foundation. The circumstances outlined in this case are a composite of real-world corporate situations, while the characters are fictional and loosely based on real professionals.

[2] "Major Depression," National Institute of Mental Health (NIMH), January 2022, https://www.nimh.nih.gov/health/statistics/major-depression (accessed Jul. 13, 2022).

This public-sourced, fictionalized case was prepared by Adriana Krasniansky (MTS '21, Harvard University) under the guidance of Mary C. Gentile, Creator and Director, *Giving Voice to Values*, and former Richard M. Waitzer Bicentennial Professor of Ethics, Darden School of Business. It was written as a basis for class discussion rather than to illustrate effective or ineffective handling of an administrative situation.

shared that Catalise's CEO was pushing to expedite Catalisten's launch time line, in order to ensure that Catalisten did not fall behind any of its competitors, which leadership believed had similar products in development.

A few days later, Taylor had a video call with Eduardo, the lead AI engineer on Catalisten's development team. The two reviewed Catalisten's production time line. At the CEO's request, Eduardo had moved Catalisten's launch date to March 1, three weeks away. Eduardo explained that the team was in its testing and validation phase, testing the performance of the software's AI model; however, some tests were being abbreviated to fit within the expedited launch time line. To help get up to speed, Taylor asked to review the completed test results, while Eduardo and the team pushed forward with their remaining work.

As Taylor reviewed the test results, she noted that Catalisten's AI software misclassified female voice samples as having MDD in 19% of test cases, but misclassified male voice samples as having MDD in only 2.1% of test cases. Taylor was concerned; false-positive diagnoses could lead to unnecessary medical treatments and costs for patients. This risk was even more pronounced for women; early research suggested that women were more likely to be overdiagnosed with MDD and overprescribed psychotropic medication.[3]

The next day, Taylor called Eduardo and explained her findings. Eduardo sighed. If the team were to address the false-positive rate for women, he explained, it would add at least six weeks to Catalisten's launch time line, upsetting the CEO and possibly endangering the company's competitive advantage. Eduardo told Taylor to let the issue go. "You've given me a lot to think about, Eduardo. Let me get back to you," Taylor said. She exited the meeting and went back to review her notes.

Background: AI, Natural Language Processing, and Medical Diagnosis

First developed in the 1950s, AI was a computer science field focused on designing computer systems and machinery that could perform tasks as well as, or better than, humans. By 2022, the field had advanced to the point where everyday technologies used AI to accomplish tasks such as language processing (e.g., translating movies), image recognition (tagging photos), and situational planning (mapping out driving routes).

There were many technical subfields within AI, one of which was natural language processing (NLP). NLP focused on developing software that could interpret human language. NLP software algorithms took input data from speech or text and analyzed it by comparing it to conventional patterns in grammar, semantics, and tone. From there, the analyzed data could be output as closed captioning, translated in real time into another language, or used to trigger a software command—such as a voice assistant interpreting human speech to set a timer or share the weather forecast. Often, NLP algorithms were used in conjunction with other AI algorithms to complete a complex string of tasks based on a single voice or text input.[4]

By analyzing speech and text patterns, NLP algorithms could also be used to inform diagnoses of mental health and neurological conditions. Everyday conversations included key cues about an individual's mental and neurological states, including speech complexity, semantic coherence (i.e., using words in the correct context), and mood valence (emotional tone of voice). When a person's free speech was run through a specifically trained

[3] Amaia Bacigalupe and Unai Martín, "Gender Inequalities in Depression/Anxiety and the Consumption of Psychotropic Drugs: Are We Medicalising Women's Mental Health?," *Scandinavian Journal of Public Health* 49, no 3 (2021), 317–24, https://doi.org/10.1177/1403494820944736; Lena Thunander Sundbom, Kerstin Bingefors, Kerstin Hedborg, and Dag Isacson, "Are Men Under-Treated and Women Over-Treated with Antidepressants? Findings from a Cross-Sectional Survey in Sweden," *BJPsych Bulletin* 41, no. 3 (2018): 145–50, https://doi.org/10.1192/pb.bp.116.054270 (both accessed Jul. 12, 2022).

[4] Dibyendu Banerjee, *Natural Language Processing (NLP) Simplified: A Step-by-Step Guide* (Data Science Foundation, 2020), https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide (accessed Jul. 12, 2022).

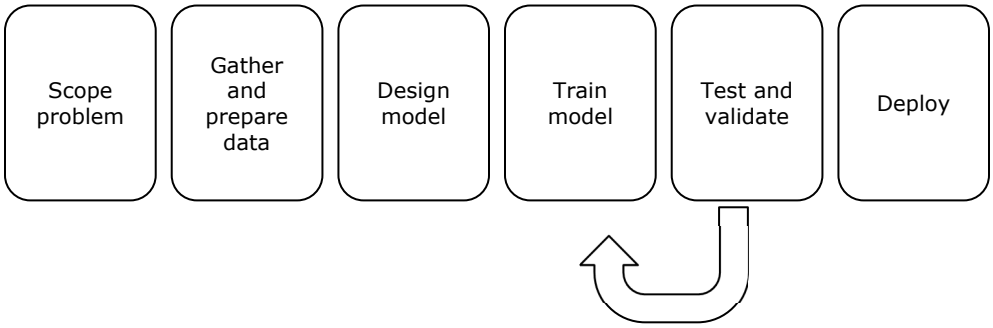
NLP algorithm, the software could calculate a statistical probability of that person experiencing a certain mental or neurological condition, based on a weighted combination of these cues.

In 2022, NLP software was used in the diagnosis and management of conditions including MDD, postpartum depression, post-traumatic stress disorder, schizophrenia, psychosis, suicidal ideation, mild cognitive impairment, and dementia.[5] Because NLP algorithms refined their calculations after each new exposure to data, these diagnostic tools could, over time, outperform doctors' classifications.[6] NLP tools could be found in hospitals and doctors' offices, where they interpreted speech and text from doctor's appointments, electronic health record notes, or the patient's bedside. They were also used in outpatient settings, analyzing language from patient portals, medical apps, social media forums, and voice assistants.

AI Product Development

Taylor and the Catalise team were using a conventional AI development process to build out Catalisten's NLP software. A conventional AI product development process followed the workflow in **Figure 1** and was overseen by a senior product manager.[7]

Figure 1. AI product development workflow.



Source: Created by author based on "The Product Management for AI & Data Science Course 2021," Udemy, 2021, https://www.udemy.com/course/the-product-management-for-data-science-ai-course/ (accessed Jul. 12, 2022).

The first step in this development flow was to clearly define a problem for the AI product to solve. At this stage, the product manager would oversee a team of user experience (UX) researchers, who would conduct interviews with prospective users to learn more about their daily lives and the challenges they faced. Based on these interviews and other research exercises, UX researchers would prepare a specific problem statement for the AI product to solve, which would then be used to guide the team in the following stages of development.

Next, the team's data scientists would begin gathering data to train the product's AI algorithm(s), collectively known as the AI model. Data scientists would look for data sets that most clearly matched situations specified within the problem statement. Data scientists also had to be careful to select data sets that were not

[5] Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen, "Natural Language Processing in Mental Health Applications Using Non-Clinical Texts," *Natural Language Engineering* 23, no. 5 (2017): 649–85, https://doi.org/10.1017/s1351324916000383; Gillinder Bedi, Facundo Carrillo, Guillermo A. Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B. Mota, Sidarta Ribeiro, Daniel C. Javitt, Mauro Copelli, and Cheryl M. Corcoran, "Automated Analysis of Free Speech Predicts Psychosis Onset in High-Risk Youths," *Npj Schizophrenia* 1, no. 1 (2015): 15030, https://doi.org/10.1038/npjschz.2015.30; Daniela Beltrami, Gloria Gagliardi, Rema Rossini Favretti, Enrico Ghidoni, Fabio Tamburini, and Laura Calzà, "Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline?," *Frontiers in Aging Neuroscience* 10 (2018), https://doi.org/10.3389/fnagi.2018.00369 (all accessed Jul. 12, 2022).

[6] https://doi.org/10.1038/npjschz.2015.30.

[7] Rahul Parundekar, "The Essential Guide to Creating an AI Product," *Towards Data Science* (blog), Medium, March 13, 2020, https://towardsdatascience.com/the-essential-guide-to-creating-an-ai-product-in-2020-543169a48bd (accessed Jul. 12, 2022).

skewed by classes like race (e.g., including more data from white individuals than Black individuals) and gender (e.g., including more data from men than women). After a set was selected, data scientists would "clean" the data, meaning they would remove any incomplete or mis-input points. Then they would "annotate" it, or review data points and selectively remove some data in order to balance class distribution within the data set.

While data scientists prepared the data set, AI engineers would design and write the code for the AI model to operate. Once a model was coded and its data set prepared, AI engineers would "train" the model. Training involved letting the AI model explore the cleaned and annotated data set to identify patterns. Based on the patterns it observed, the AI model would begin to predict which data features predicted certain outcomes. As the model became more confident in its predictions, it would start to recommend decisions for different data points based on the features it recognized.

Once an AI model was sufficiently trained on the cleaned and annotated data, the team's AI engineers would begin to "test" the model by introducing it to data sets that more closely resembled the messy and complex data found in the real world. Every time the model stumbled (i.e., its predictions dropped in accuracy), the team would regroup and prepare new training data for the model to learn from. This testing and validation cycle would continue until the team determined that the model was ready for deployment into real-world settings.

Evaluating Catalisten: Data Bias and Model Accuracy

When Taylor joined the Catalisten team, Eduardo and his engineers were actively testing and validating the product's AI model. And, as Eduardo mentioned, the team was abbreviating this step as part of Catalisten's expedited launch.

Taylor wondered which points in the development process had allowed for the discrepancy between female and male false-positive rates. She pored over the team notes from all phases of Catalisten's development process and found two important inflection points. First, in the data preparation phase, Catalisten's data scientists reported difficulty compiling a training data set that presented a balanced number of female and male patients. All the patient data sources to which the Catalisten team had access presented significantly higher percentages of men than women. And, because Catalisten's team was separated from colleagues and research partners to avoid competitive leaks, its data scientists could not reach out to health systems or research partners for better data sets.

Buried in the team's workflow chat logs, Taylor found a note from the team's head data scientist confirming a gender bias in the data set, while also indicating that the women represented had a higher-than-average rate of MDD. The head data scientist warned that, if trained on these data, Catalisten's AI model would be more likely to overdiagnose women with MDD. However, given timing constraints, Eduardo—who worked as the interim Catalisten product lead before Taylor's arrival—instructed the data science team to continue with the data set despite its bias. If Catalisten's model did "learn" this bias during its training, Eduardo promised that it would be addressed in the testing and validation stages.

Taylor scrolled further down the chat logs to see what happened next. When Catalisten's model entered its testing and validation phase, engineers confirmed that it presented a higher rate of false positives for women than for men. This finding sparked a heated conversation among engineering team members in the chat log regarding *precision* and *recall*, two different evaluative metrics for AI model accuracy. Precision answered the question, "When the model makes a prediction, how likely is that prediction to be correct?" It was calculated by dividing the model's number of true positives by all positives, and it was higher when the number of false positives was low. Recall, on the other hand, answered the question, "How good is a model at identifying actual

occurrences of objects in the data?" It was calculated by dividing the model's true positives by its true positives plus false negatives, and it was higher when the number of false negatives was low.[8]

Precision and recall metrics were often in tension, and AI product teams had to evaluate the impact of each metric on model performance. If the Catalisten team were to prioritize precision, the AI model would be more conservative in its MDD diagnoses; it would have a very low rate of false positives but would also miss some MDD cases (i.e., present a higher number of false negatives). If the Catalisten team were to prioritize recall instead, the AI model would diagnose a greater number of MDD cases but would also falsely diagnose some individuals (i.e., present a higher number of false positives).

Catalisten's chat logs captured the back-and-forth between engineers as they discussed whether to prioritize the product's precision or recall. The product had a high recall, correctly diagnosing all MDD cases but also diagnosing some individuals with MDD who did not have the condition. A few engineers proposed that high recall was a good thing, ensuring that everyone who needed MDD treatment received it. High recall also presented a strong business case for Catalisten, since mental health clinic customers would certainly see an uptick in diagnoses made—driving more revenue from patient procedures and prescriptions. Other engineers pointed out that Catalisten's MDD overdiagnoses, especially for women, would lead to patients being prescribed unnecessary medications, which could negatively affect their emotional stability, social interactions, and even legal rights.

The chat logs showed that Eduardo had made the final team decision. High recall was a positive feature for Catalisten's customers, Eduardo noted, and lowering Catalisten's false-positive rate for women would require more testing and validation—pushing back the launch time line. "Let's not draw any more attention to the 19% false-positive rate for women," he wrote. "Any fallout effect from the false-positive discrepancy isn't our problem. It's in the hands of our customers who write the prescriptions—they know that all technologies carry risks. Our job is to get this product to market as fast as we can."

Decisions Moving Forward

Taylor took a walk to process everything she had learned since joining the Catalisten team. She reflected on her conversation with Eduardo and sympathized with his experience as Catalisten's interim lead. Start-ups were notorious for promoting a "move fast" mentality, and Catalise was no exception in prioritizing speed to market. Taylor understood that Eduardo had sacrificed important steps to address the gender bias in training data and chosen not to correct the model's gender bias during the testing and validation phases because he faced outsized pressure from leadership to expedite the Catalisten launch.

Yet Taylor also knew that Catalise's company culture prioritized going above and beyond for patients. If Catalisten became the groundbreaking success that leadership hoped it would be, it was only a matter of time before the public uncovered its high false-positive diagnosis rate for women. So, if Catalisten launched without addressing the issue, not only would it risk female patients' health, but it would also jeopardize the company's reputation and that of its employees. Taylor didn't want her product management career or her teammates' careers to be tarnished by a rushed product launch. She needed to counteract Eduardo's decisions and convince Darnell to delay launch.

Imagine that you are Taylor. How would you go about convincing Darnell, your chief product officer, to delay Catalisten's launch in order to address its gender bias?

[8] Adapted from "The Product Management for AI & Data Science Course 2021," Udemy, 2021, https://www.udemy.com/course/the-product-management-for-data-science-ai-course/ (accessed Jul. 12, 2022).