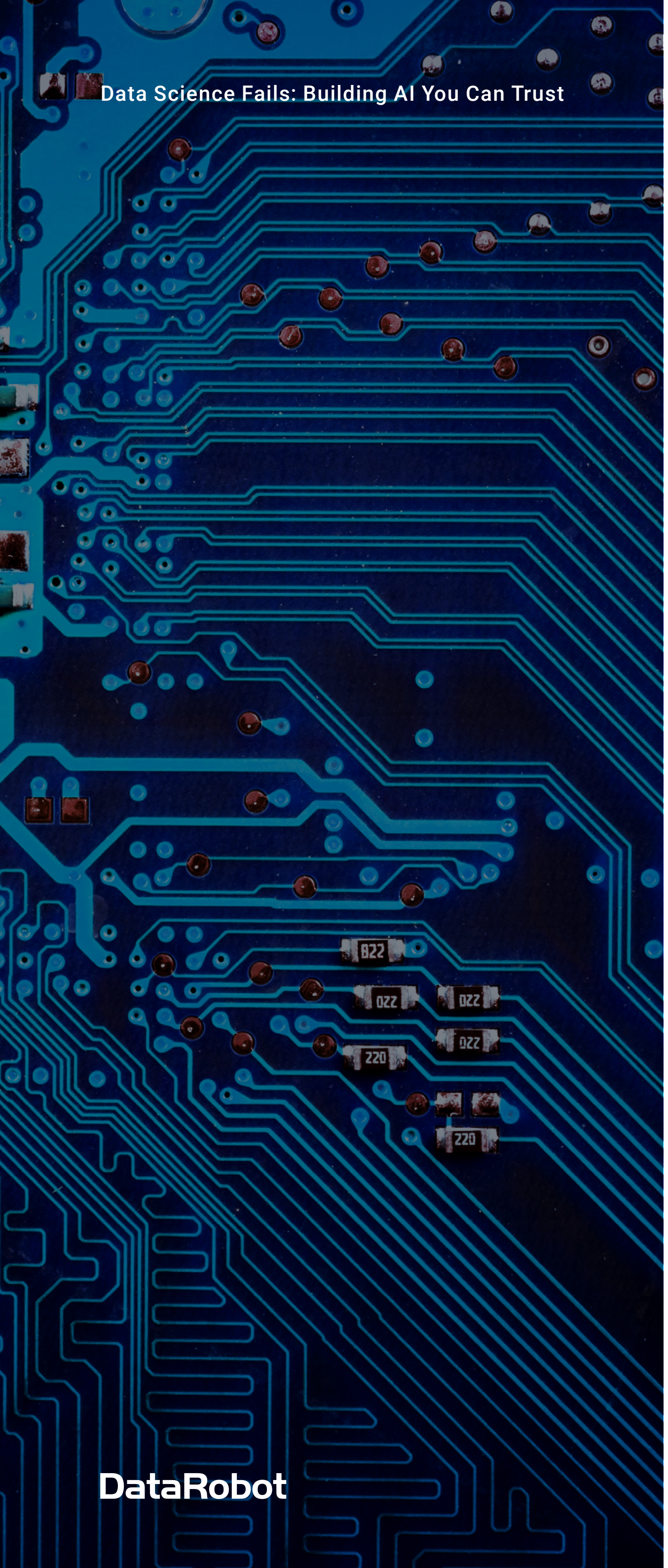
Data Science Fails: Building AI You Can Trust

DataRobot



Trusting Your AI

It's hard to read technology news these days without coming across a headline pointing to the game-changing potential of artificial intelligence (AI) and machine learning solutions. And with good reason. After all, the use cases for AI are well-documented and individuals and businesses worldwide are eager to put the technology to work to solve numerous challenges and issues.

Before AI can be put to use, however, business and other organizations must first be able to trust any AI solutions they decide to implement. **Trusting AI means carefully considering the solution's reliability and whether the AI is being swayed by human bias.** Following the proper data science guidelines can help organizations to trust their AI solutions and have confidence that their models are working correctly and effectively. Additionally, these practices can make AI more transparent, instead of a black box in which the decision-making processes behind AI algorithms are hidden from users.

So how can organizations build AI systems that are trustworthy, follow best data science practices, and reflect their core values? Sometimes to understand how to get data science practices right, it's important to understand how and why data science sometimes goes wrong. This Data Science Fails report — based on insights from DataRobot's Vice President for AI Strategy, Colin Priest — outlines several important lessons on how organizations can implement AI successfully. Using eight examples from players who were unsuccessful in their initial efforts, we can all learn from these mistakes on how to prevent AI bias.

There's No Such Thing As A Free Lunch

The *no free lunch* theorem states, in short, that no algorithm can be equally good at learning everything, which means that you can't know in advance which algorithm will work best on your data. Despite this well-established theorem, it is common practice for data scientists to rely on only a limited number of modeling methods.

In August 2018, *Nature* published the article "Deep learning of aftershock patterns following large earthquakes." Using a training dataset containing more than 130,000 mainshock — aftershock pairs, the authors trained a deep learning algorithm to "identify a static-stress-based criterion that forecasts aftershock locations without prior assumptions about fault orientation."

However, the article and subsequent hype should have raised a couple of red flags:

1

First, forecasting seismic activity is challenging. The accuracy looks too good to be true. The paper uses only a single machine learning algorithm, applying only one architecture.

2

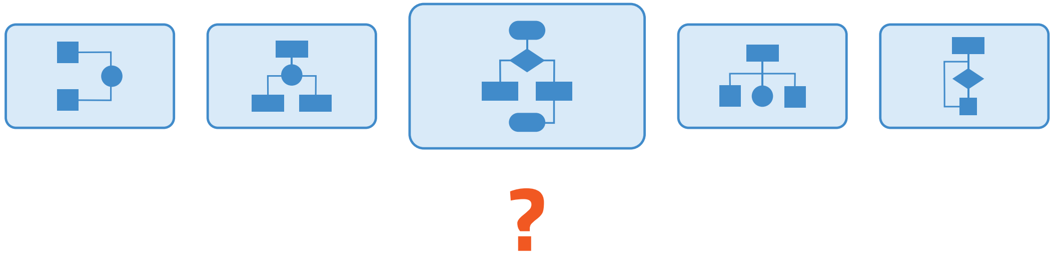
Second, at the same time that the *Nature* article was published, industry analysts were commenting on the hype associated with certain technologies with analysts listing "Deep Neural Networks" as being at the peak of the hype cycle.

A year later, *Nature* published a follow-up article using the same data but written by different authors. This new article compared the original model against a simple linear logistic regression model, concluding that the simpler model "provides comparable or better accuracy." The new authors further concluded that "deep learning does not offer new insights or better accuracy in predicting aftershock patterns."

Data Science Lesson: **Regulate Your AI Bias**

HUMANS CAN BE BIASED IN CHOOSING ALGORITHMS

You may have your favorites, or you may be excited to try the coolest and latest algorithms. But to avoid an algorithm bias, step aside and let competition between a champion and challenger model decide which method is superior. A lack of diversity in model building usually leads to suboptimal results.



Be Careful What You Wish For

In a *Science* journal article, "Dissecting racial bias in an algorithm used to manage the health of populations," published in October 2019, the authors describe a data science failure. When a health system anticipates that a patient will have extraordinarily complex and intensive future healthcare needs, that patient is enrolled in a ''care management'' program, which provides considerable additional resources: greater attention from trained providers and help with coordination of care. Some healthcare systems use an algorithm to determine a patient's "commercial risk score," which, in turn, is used to select which patients are given access to care management.

Researchers accessed detailed data on primary care patients in a large teaching hospital and decided to use that data to test the behavior of a widely used algorithm to measure differences in outcomes between the self-identified race of patients.

At first, the research only looked at clinical data. When the researchers expanded their analysis to other patient information, they discovered that one of the best predictors for length of stay was the patients' zip code. The zip codes that correlated to longer hospital stays were in poor and predominantly African-American neighborhoods. Then, when they mapped commercial risk scores versus the number of active chronic conditions, split by race, they discovered that African-American patients with the same number of chronic health problems received lower commercial risk scores.

This bias arose because the algorithm predicts healthcare costs rather than illness, but due to entrenched historical disadvantages, less money had been spent in the past on care for African-American patients, and the algorithm unintentionally learned to replicate that outcome.

Data Science Lesson: **Adopt Best Practices for AI**

The problems in the case study arose due to incorrect goals, perverse incentives, and inadequate model governance processes. The lessons to be learned are:

Define your organization's values for AI ethics and publish those values as internal policy guidelines. These guidelines will inform the development of new AIs and the design of appropriate governance processes.

When practical, build your own AIs, avoid using third-party-sourced, black-box AIs that may not share your values, and insist that your AI's decisions are explainable and justifiable. For use cases with a significant risk of harm, such as in healthcare, apply rigorous governance both before and after deploying AI systems. If you must use a third-party system, ask the vendor to provide documentation of how the AI was trained, what goal it optimises, and how the algorithm for tested for unfair bias.

Be careful what you wish for, you might just receive it. AIs can be ruthless in optimizing the goals you give them. For AI you can trust, apply best practices for AI governance, carefully defining your goals, and test the AIs behavior using the latest generation of AIs that provide human-friendly explanations.

Watch Out for Typos

In November 2015, a study was published that concluded that:

- Family religious identification decreases children's altruistic behaviors
- Religiousness predicts parent-reported child sensitivity to injustices and empathy
- Children from religious households are harsher in their punitive tendencies

The research was carried out with children in six countries (Canada, China, Jordan, Turkey, South Africa, and the United States), and included 510 Muslim, 280 Christian, and 323 nonreligious children. It was the first research to take a large-scale look at how religion and moral behavior interact in children from across the globe.

Researchers immediately raised questions about the study. A University of Oregon psychologist told *Science Magazine* in 2015 that he was confused by the results as they didn't match previous research that, taken as a whole, found no overall effect of religion on adults faced with these kinds of moral tests. He requested that the authors share their data so that he could understand why the paper obtained such different results versus previous research.

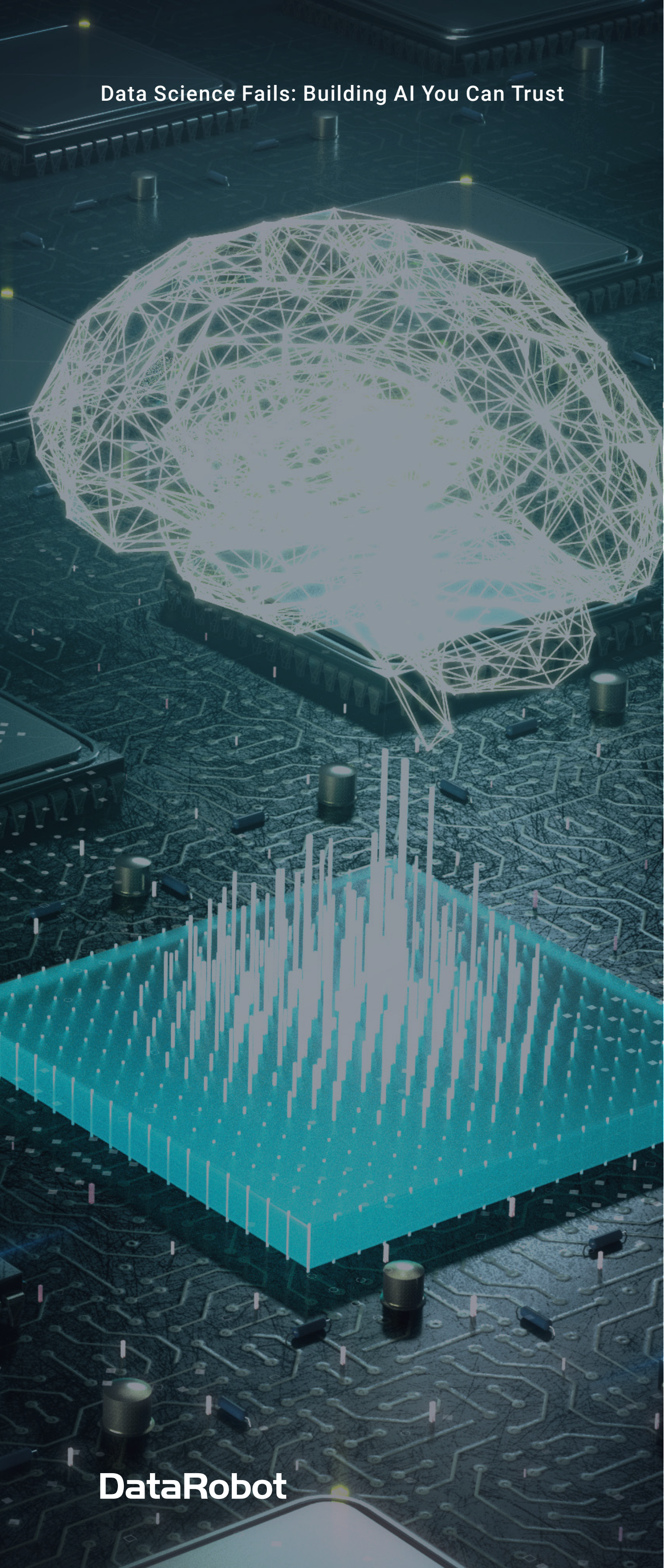
The confusion came down to a simple typo. When coding in their results, they used numbers to represent each country — 1 for the US, 2 for Canada, and so on. After correcting the mistake, the researchers discovered that country of origin, rather than religious affiliation, was the primary predictor of several of the outcomes.

A simple typo reversed the conclusion of the research.

Data Science Lesson: **Minimize Opportunities for Human Error**

TYPOS OCCUR EASILY AND ARE AN INEVITABLE RESULT OF MANUAL PROCESSES

The solution is to automate these processes with well-tested, widely used tools and libraries. In data science, the solution is automated machine learning that reduces the dependence on manual scripting and has guardrails to identify possible errors.



Change Is the Only Constant

In 2018, ABC Investments (not their real name) launched a fund that utilizes machine learning to identify sources of potential returns. The launch material promised that machines would “forecast market moves more accurately” and apply “an elegant approach with great return potential.” In short, investors in the fund were placing a multimillion-dollar bet that algorithms would be more effective at figuring out the complex world of discretionary investing than a human portfolio manager.

In the fine print at the bottom of the announcement of the AI fund launch, ABC included the standard investment product caveat: “Past performance is not a guide to future results.” This caveat was to prove prophetic. Over the next twelve months, the fund showed poor returns, underperforming the share market. An ABC spokesperson explained the underperformance as “the result of challenging markets and changing behavior of so-called equity factors.” This was a surprising excuse to make, not only due to the promise at fund launch that machines would be more accurate at timing the market, but also because it is well known that equity factors change all the time; they are not stable. Just like many others who have tried and failed, ABC was unable to use AI to predict equity markets with sufficient accuracy to consistently outperform an index.

So why is equity market price prediction the wrong use case for AI?
It's not enough for an algorithm to have worked in the past. Algorithms that were trained on data sourced from a time period during which only a single investment regime occurred will need to be retrained to learn the new market paradigms. And unless the algorithms recognize the regime change, they will be slow to update their investment strategies.

Data Science Lesson: **Constantly Monitor Your AI**

A best practice internal AI governance and risk management process will include plans that cover the following general principles:

Choose the right use cases for AI. Most modern AIs rely on pattern recognition, using machine learning techniques to find those patterns. The most successful AI use cases incorporate past patterns can be reliably extended into the future.

AIs can be overconfident, just like humans. Follow best practices to ensure AI humility. Ensure that your AI warns you when it is about to make a decision using data that is outside the range of its training data, and set up AI processes so that the system automatically triages difficult decisions and edge cases to humans.

Managing an AI can be similar to managing human staff. Just as you would manage human experts using performance indicators and regular training to update their skills, you should not deploy an AI and leave it to run without performance tracking and updates. Best practice is to use MLOps systems to proactively warn when live data is dissimilar to training data or when accuracy is deteriorating.

If It Looks Too Good To Be True

Target leakage, sometimes called data leakage, happens when you train your algorithm on a dataset that includes information that would not be available at the time of prediction and then apply that model to data you collect in the future. Since it already knows the actual outcomes, the model's results will be unrealistically accurate for the training data, like bringing an answer sheet into an exam.

For example, a research paper published in *Biomedical Signal Processing and Control* in January 2020 boasts of achieving "100% [congestive heart failure] detection accuracy" using a new deep learning model that predicts congestive heart failure "on the basis of one raw electrocardiogram (ECG) heartbeat only." Like the earthquake model, the news seemed too good to be true. And it was.

The first warning sign was achieving 100% accuracy. Perfect accuracy is usually an indication of a problem. Either the outcome is trivially easy to predict, or you've accidentally used data that exploits knowledge of the outcomes.

The second warning sign was achieving accuracy from a single heartbeat. That isn't much data from which to infer a health outcome. ECG data is notoriously noisy, and to compound the problem, the data used in the study used only "lead 1" values, (i.e., sourced from only a single electrical lead).

The 100% accuracy headline, used in the abstract and news articles, is misleading, since it was calculated on the training data. Whenever you measure the accuracy on the training data, you overestimate the accuracy. It is quite common for a model to overtrain, to memorize the training data and not perform as well on new data. In this case, the accuracy on test data was only 97.8%.

After delving more deeply into the data, we discovered problems with the experimental design. Even though there are half a million rows of data, there are only 33 subjects. That means there are only 33 independent outcomes, not enough to trust that the model has generalized to work for most of the population. The risk is that, instead of predicting congestive heart failure, the model learned to identify individual subjects, as a cheating shortcut to predicting the outcome.

Of the 33 subjects in the study

18 were normal healthy subjects

while 15 subjects had severe congestive heart failure

However, the healthy subjects came from a different database than unhealthy subjects. The two databases were not directly comparable. Furthermore, the two data sources stored the heartbeat data using different time frequencies, requiring the high frequency data to be downsampled to be comparable to lower frequency data, creating pre-processing artifacts in the data. It is quite likely that the algorithm is cheating by using the characteristics of the transformed data source, rather than the attributes of heartbeats from healthy versus unhealthy hearts.

Data Science Lesson: **Verify Your AI's Results**

The tools used to build machine learning and AI models are becoming easier to access and use but are still vulnerable to subtle errors that can easily lead you to the wrong conclusions if you're not careful. Be suspicious of your model if the results seem too good to be true. Follow best practices to reduce the chance of errors:

Measure accuracy only on out-of-sample data, such as the validation data or the holdout data, never from the training data.

Partition your data carefully. Don't allow data from the same subject (whether that be a person or a geological fault line) to be allocated across multiple partitions.

Check for target leakage. When possible, use tools that have guardrails that detect potential target leakage.

Employ common sense behavior and check which input features are the most important. For the most important input features, determine which values lead to yes or no decisions, or high or low predictions.

Take extra care with data that spans multiple periods. It is very easy for partial target leakage to creep in via your feature engineering or hyperparameter tuning. When possible, use tools that are capable of time-aware model training.

Ignoring Business Rules and Expertise

In 2011, IBM Watson wowed the tech industry and cemented a place in pop culture with its win against two of Jeopardy's greatest ever champions. Ken Jennings and Brad Rutter were the best players the show had produced over its decades-long run, the former boasting the longest unbeaten record and the latter earning the highest total prize pool.

IBM researchers trained Watson by giving it thousands of Jeopardy clues and possible responses that had been manually labelled as correct or incorrect. From this data, Watson discovered patterns and made a model for how to get from an input (a clue) to an output (a correct response). While Watson was not able to solve every question correctly, it proved itself capable of outperforming the best human players.

With a convincing Jeopardy win behind it, Watson was looking for a new challenge. By turning its natural language processing abilities to medicine, Watson could read patients' health records as well as textbooks, peer-reviewed journal articles, and lists of approved drugs. Maybe, with access to all this data, Watson might find patterns and solutions that no human could ever spot.

> In October 2013, IBM announced that The University of Texas MD Anderson Cancer Center would use Watson "for its mission to eradicate cancer."

However, rather than playing to Watson’s strengths in natural language processing and linking questions to answers from a text database, Watson was only trained on a small number of “synthetic” cancer cases, or hypothetical patients. It was not trained on real patient data. IBM’s internal records showed that the number of training cases varied for different types of cancer, from only 635 cases for lung cancer, down to only 106 training cases for ovarian cancer. The results showed that these training cases were not enough for Watson to learn how to make accurate decisions.

Of course, healthcare is not like Jeopardy. If you make a mistake competing in Jeopardy you can still win, but if you make a mistake in healthcare, someone's life could be at risk. Internal IBM documents show that Watson often gave erroneous cancer treatment advice and that company medical specialists and customers identified “multiple examples of unsafe and incorrect treatment recommendations.”

Data Science Lesson: **Understand Your AI's Decision-Making**

When AI is being asked to make decisions with significant consequences, such as life-and-death healthcare recommendations, it needs to be trustworthy. The more consequential your AIs decisions, the greater the care you must apply. For AI you can trust, follow best data science practices:

Don't make an AI learn something you already know. Don't discard prior knowledge. Feature engineer the business rules and expertise into the algorithm.

Use enough data for true data relationships to be distinguishable from spurious correlations. You can't get accurate models on difficult problems using only 106 training examples. Use learning curves, which track improvements in accuracy as you add more training examples.

Don't do big projects all at once. AIs are better suited to narrow tasks than many diverse tasks. Break up large projects into smaller, more achievable use cases. Solve one at a time, one AI per task, leaving the most difficult or time-consuming use cases until last.

Before deploying a new AI, have its behavior reviewed and signed off by a subject matter expert. Ask them to point out any AI behavior that doesn't make sense, especially input features that don't seem to have a plausible mechanism for being related to the outcome.

the deselected mirror modifier object
modifier_ob
) # modifier ob is the active ob

Fake News, Fake Data

In January 2018, Twitter admitted that more than 50,000 Russia-linked accounts used its service to post automated material about the 2016 U.S. election. The posts had reached at least 677,775 Americans. In response, Twitter removed 50,258 accounts and passed their details to investigators. Similarly, an investigation into fake news published on Facebook found that fake news on political topics reached 158 million people.

Social media companies have hired thousands of employees to prevent the spread of fake news on their platforms. Yet, with so much disinformation occurring at such a scale, it is impossible for humans to manually detect and correct all fake news. And now that disinformation can be automated by using AI, the task of manual detection is looking even more hopeless. To guard against this threat, researchers have begun training AIs to automatically detect fake news. While the research field is young, there has been progress with promising results.

One approach detects fake news using stylometry-based provenance, (i.e., tracing a text's writing style back to its producing source and determining whether that source is malicious). The stylometry-based approach assumes that fake news can be identified solely by determining the source that generated the text. Researchers report achieving up to 71% accuracy with a stylometry-based approach, depending on the dataset used.

Another approach trains against the FEVER (Fact Extraction and VERification) dataset, which consists of 185,445 "claims" generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the text from which they are derived. Rigorous fact verification, used in manual processes for identifying fake news, requires validating a claim against reliably sourced evidence. However, for the FEVER dataset, claim-only classifiers (which don't validate against the evidence) perform competitively against the top evidence-aware models. Researchers report achieving predictive accuracy of up to 61% on the FEVER dataset.

Accuracy percentages of 71% and 61% may not be perfect, but they do signal progress in the war against fake news. However, recently published research by an MIT team shows that we shouldn't take these accuracy rates at face value. The problem is that fake news detection developers had chosen an easy-to-beat benchmark.

Data Science Lesson: **Be Careful What You Teach AI**

It seems that AI is better at creating fake news than identifying it. But the lesson learned from the case study on detecting fake news can be applied to any AI project. It is essential that you apply critical thinking skills to training and evaluating your AIs:

Beware of using fake data to train your AI. Fake data, whether simulated or crowdsourced by humans, is usually not the same as real data. Machine learning algorithms will try to cheat, to learn to identify the artifacts of the data creation process instead of the characteristics of real life.

Beware of using proxies for the outcome that you wish to predict or decide. Those proxies may not safely align with your intended outcome.

Insist that your AIs provide human-friendly explanations for how they are working and why they made their decisions. Then check whether those explanations are showing whether the AI is finding true-to-life patterns, or just cheating.

The Transparency Sweet Spot

In 2013, Stanford University professor Clifford Nass received complaints from students enrolled in one of the two course sections of his technological interface course. The students believed that their section of the course unfairly received lower grades than students in the other section. When Professor Nass investigated, he discovered that the complaints were justified. Two different teaching assistants had marked the two student sections, and students with similar answers had received different grades.

Using his computer science skills, Professor Nass devised an algorithm that would correct for the human bias, giving higher grades to the students whose exams had been marked by the less generous exam marker. In the spirit of full transparency, he shared the full details of the algorithm. Yet, some students were even angrier than before.

In 2015, René Kizilcec, a PhD student who had worked with Professor Nass, decided to conduct a research study to look at the effects of grading transparency on student trust. In Kizilcec's study, 103 students submitted essays for peer grading. Internally, the grading process returned two marks: a grade that represented an average peer grade and a "computed" grade which was the product of an algorithm that adjusted for bias.

Students were randomly given one of three levels of transparency about how their final grade was calculated:

Low Transparency

Students received the computed grade.

Medium Transparency

Students received the computed grade accompanied by a paragraph explaining how the grade had been calculated, why adjustments had been made, and naming the type of algorithm used.

High Transparency

Students received the computed grade accompanied by a paragraph explaining how the grade had been calculated, why adjustments had been made, and naming the type of algorithm used. These students also received their raw peer-graded scores and saw how these scores were each precisely adjusted by the algorithm to arrive at the final grade.

Students in each of the three groups were asked to rate their trust in the process. The final results were published in 2016 in the paper, "How Much Information? Effects of Transparency on Trust in an Algorithmic Interface." Using the data from the study, Kizilcec arrived at three key conclusions:

- Individuals whose expectations were not met (by receiving a lower grade than expected) trusted the system less, unless the grading algorithm was made more transparent through explanation.
- However, providing too much information eroded this trust.
- Attitudes of individuals whose expectations were met did not vary with transparency.

The observations, while disproving the common perception that greater transparency is always better, described intuitively reasonable behaviors. People distrust an algorithm when it deviates from their expectations and appears to be disadvantageous to them. Providing information can sometimes help to align human expectations with eventual outcomes, but too much transparency can have a negative effect.

Kizilcec concluded, "Designing for trust requires balanced interface transparency — not too little and not too much."

Data Science Lesson: **Find The Right Balance of AI Transparency**

Provide explanations for algorithmic decisions to key stakeholders. Professor Nass could have prevented the student complaints via a strategy of optimal transparency. An approach of no disclosure and no transparency leads to inadequate AI governance, AI behavior that is inconsistent with business rules and core values, and dissatisfied stakeholders.

Don't overcommunicate how an algorithm works. At the same time, a strategy of full transparency results in confusion and suspicion, disgruntled stakeholders, loss of intellectual property, and risks gaming of the system, (e.g., fraudsters discovering tricks to avoid detection and exploit a system's vulnerabilities).

Use human-friendly heuristics to explain algorithmic decisions. Optimal transparency uses heuristics to create human-friendly explanations of the reasons for a decision, which is crucial for decisions that have adverse outcomes for stakeholders and seemingly counterintuitive decisions that defy stakeholder expectations.

The Advent of Automated Machine Learning and Enterprise AI

Automated machine learning, which was invented by DataRobot, can create Enterprise AI applications to address many of the challenges described above and make others more manageable.

Enterprise AI applications:

Reduces the occurrence of human errors by **automating manual tasks**, such as feature engineering, missing value imputation, hyperparameter tuning, and model comparisons. Reduces the opportunity cost of using data scientists for mundane repetitive tasks, freeing their time for more valuable human tasks, such as applying business rules and ethical values.

Provides transparency into each model's use of data, telling you not just which features in the data had the most impact to the predictive power of each model, but also explains individual predictions down to specific data features and their values.

Applies best-practice data science, as used by the world's top data scientists. **Guardrails** alert users to potential problems, such as target leakage, and block users from common data science errors, such as measuring accuracy on training data.

Provides **tools for understanding** model accuracy and making tradeoff decisions (e.g., between speed and accuracy, positive versus negative predictive value, when and where additional models may be cost justifiable).

Makes **retraining models** on new data and redeploying models into production simple, fast, and low risk.

Eliminates human bias from choice of algorithms so you can quickly evaluate and select the model best suited to your particular problem.

Makes it easier to **monitor model performance** and detect drift or performance degradation over time, alerting modelers to the need for retraining or creation of challenger models.

Conclusion

For organizations to successfully implement AI, they must first have confidence that their AI is reliable and trustworthy. As the previous examples highlighted, there are many ways for organizations to get AI wrong before they get it right.

Understanding and embracing best data science practices will be key to avoiding common AI missteps. This includes acknowledging that AI can be influenced by human biases, which can ultimately make the systems unreliable. Getting ahead of these issues is key to successfully becoming an AI-driven enterprise.

AI needs to be bias-free in order to be trustworthy. Adopters should also prepare data carefully and consider which data sources are informing AI decision-making and outcomes. Many organizations have already discovered that whoever owns your AI also owns your customer experiences, and as a result, your business. Organizations must, therefore, fully own their AI to ensure that they understand how decisions were reached and that they alone control their AI products. Governance structures must be established to ensure that AI abides by and reflects a company's values in practice to avoid the risk of reputational harm.

When you are working to help your organization become an AI-driven enterprise, don't let a data science fail (or two or more) stall your mission. Let DataRobot guide you on the path to best data science practices and help your organization become AI-driven.

DataRobot

DataRobot helps enterprises embrace artificial intelligence (AI). Invented by DataRobot, automated machine learning enables organizations to build predictive models that unlock value in data, making machine learning accessible to business analysts and allowing data scientists to accomplish more faster. With DataRobot, organizations become AI-driven and are enabled to automate processes, optimize outcomes, and extract deeper insights.

Sign up for a free trial today to find out how DataRobot can help your organization at **datarobot.com**