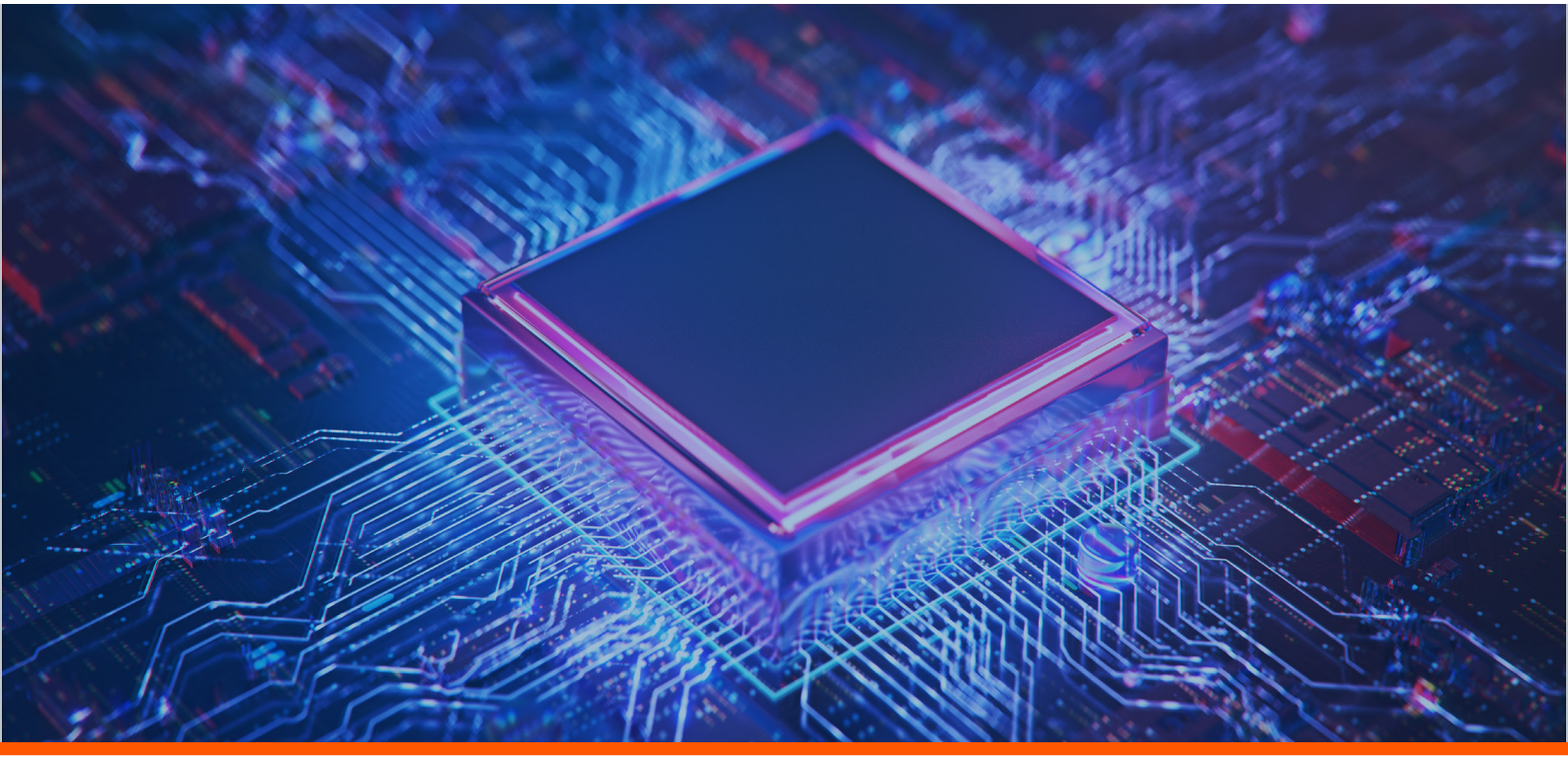


WHITE PAPER

AI Ethics

DataRobot

Contents

Introduction 1

Four Principles of AI Ethics 2

Principle 1: Ethical Purpose 2

Principle 2: Fairness 4

Principle 3: Disclosure 9

Principle 4: Governance 10

Conclusion 12

AI Ethics

Introduction

As machine learning and artificial intelligence (AI) usher in the Fourth Industrial Revolution, it seems like everyone wants to participate, and who can blame them? AI promises improved accuracy, speed, scalability, personalization, consistency and clarity. And with AI already succeeding in many industries including healthcare, finance, marketing, manufacturing, and agriculture, it's likely that your competitors are already adopting AI.

But some businesses are wavering in their decision to move forward. On the one hand, they know that they need to embrace AI innovation to remain competitive. On the other hand, they know that AI can be challenging. They've heard news stories of high profile companies making mistakes with AI, and they are worried that it may happen to them too, damaging their reputation. In regulated industries, there's the question of how to explain AI decisions to regulators and customers. Then there's the challenge of how to engage with staff so that they embrace organizational change. How do you manage AI to ensure that it follows your business rules and core values?

> "If employees have thought about proper ethical limitations of AI, they can be important guards against its misuse"
>
> **Emma Martinho-Truswell, "3 Questions About AI That Nontechnical Employees Should Be Able to Answer", Harvard Business Review Online**

You can't have a society without cooperation. You can't have cooperation without trust. Trust is a central part of all human relationships, including romantic partnerships, family life, business operations, politics, and medical practices. For example, if you don't trust your doctor, it is much harder to benefit from their professional advice. Just as human society is based upon cooperation and trust, the successful AI-driven enterprise is built upon AI you can trust. The latest generation of AI applies human values and provides human-friendly explanations for its decisions.

Ethical behavior is good for business. Consider the string of recent corporate scandals and the brand damage they have done to the businesses involved. Yet the benefits go beyond brand value. For example, in the study "Doing Well by Doing Good: The Benevolent Halo of Corporate Social Responsibility", marketing professor Alexander Chernev concluded that acts of corporate social responsibility, even when they are unrelated

to the company's core business, influence consumer perceptions of the functional performance of the company's products. The products of companies engaged in socially responsible activities are likely to be perceived as being higher quality. And the benefits of ethical behavior go beyond customer perception. The research paper "Ethics as a Risk Management Strategy", concluded that "there are compelling reasons to consider good ethical practice to be an essential part of ... risk management" and that the benefits of ethical behavior include the identification of potential risks, fraud prevention, and reduced court penalties.

Trusted AI requires that AIs apply human ethical standards. Ethics goes beyond laws and regulations - just because something is legal, doesn't mean that it is ethical. And while the details of ethical standards may vary from person to person, from organization to organization, there are general principles that apply.

> "As LinkedIn encouraged members to join conversations, it found itself in danger of creating a "rich get richer" economy in which a few creators got an increasing share of all feedback. Highly skewed distribution of feedback occurs naturally in any system that distributes content virally, but that doesn't mean it's good for creators. As a result, LinkedIn implemented changes to balance out this increasing skew and added creator-focused metrics."
>
> **Bonnie Barrilleaux, LinkedIn, "Perverse incentives in metrics: Inequality in the like economy", Stata Data Conference September 2018**

Four Principles of AI Ethics

Principle 1: Ethical Purpose

Every AI has a task and an objective. An AI's task is much like your staff's job descriptions and is defined as the decisions an AI is empowered to make. This may be which product to recommend, whether to flag a transaction as fraudulent, or what price to charge a customer. An AI's objective is much like your staff's key performance indicators and is defined as the measure that the AI must optimize. This may be maximizing sales revenue, minimizing risk, maximizing profit margin, or reducing expenses.

Most modern AIs are powered by machine learning. They learn by example, using historical data. Modern machine learning algorithms can be ruthless in optimizing their goals. They know nothing except the historical data and the objective that you set them. Just like humans, AIs are subject to perverse incentives, maybe even more so than humans. So it stands to reason that you need to carefully choose the tasks and objectives that you assign to AIs.

Firstly, consider the task you assign to an AI. Consider the comparative strengths of computers and humans. Computers are strongest at repetitive tasks, mathematics, data manipulation and parallel processing. So long as a task can be defined as a procedure, a computer can do that task over and over again, without getting tired, giving the same results

each time. Computers can manipulate numbers and data in volume much faster than any human. This does not mean that humans are obsolete. Humans are much more skilled than machines at communication and engagement, context and general knowledge, creativity and empathy. When your customers have a frustrating problem, they want to talk to a human, someone who will understand their exasperation, listen to their experience and make them feel valued as a customer, while also solving their problem. Humans are much better at common sense than computers, instantly recognizing when a decision doesn't make sense. And humans can be creative, not just with art, but creatively solving your customer's problems when the computer cannot help them because it merely follows rules for well-known problems.

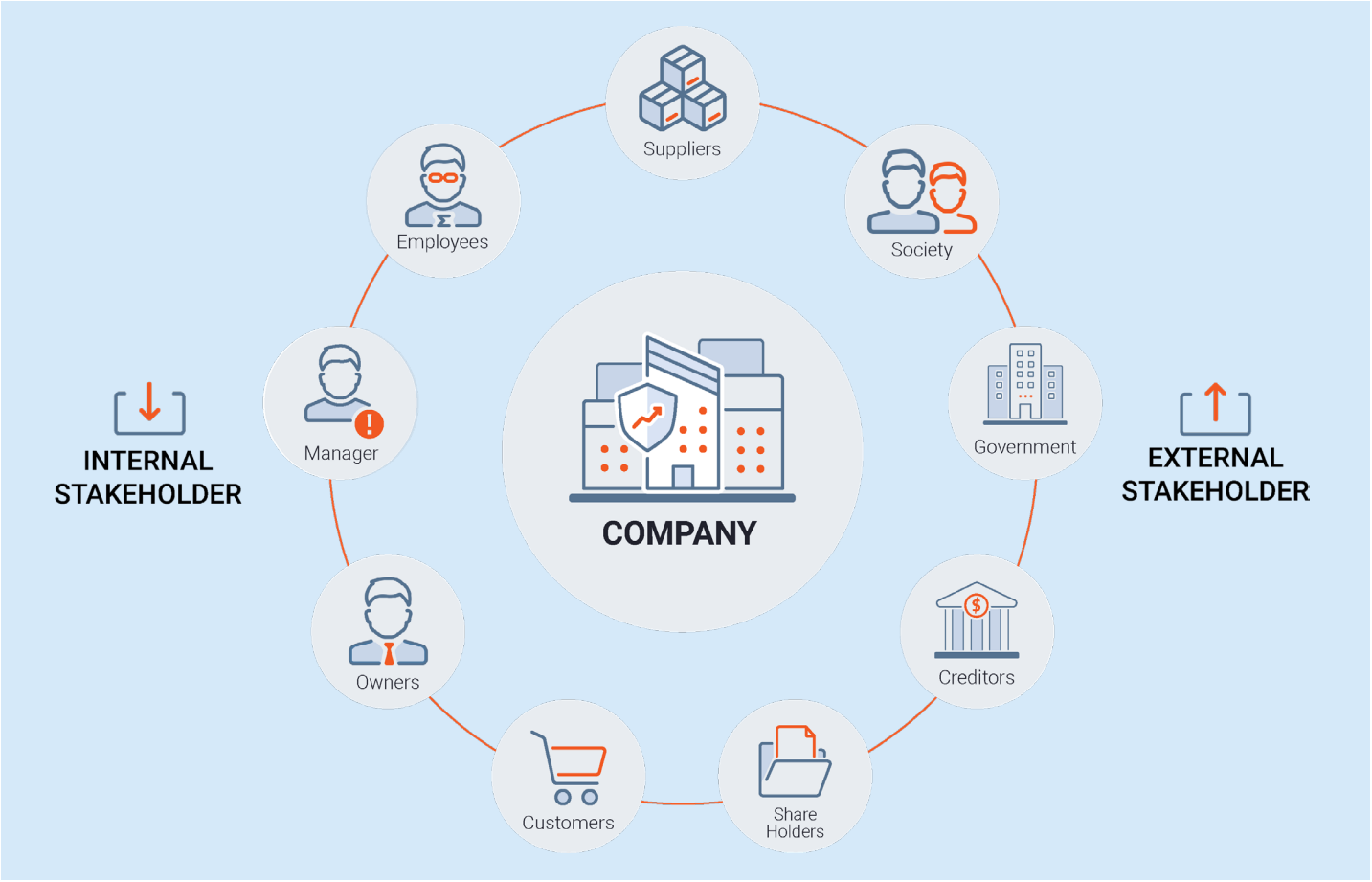
> "What if we were to reframe the situation? What if, rather than asking the traditional question—What tasks currently performed by humans will soon be done more cheaply and rapidly by machines?—we ask a new one: What new feats might people achieve if they had better thinking machines to assist them? Instead of seeing work as a zero-sum game with machines taking an ever greater share, we might see growing possibilities for employment."
>
> **Thomas Davenport and Julia Kirby, "Beyond Automation", Harvard Business Review July 2015**

Are you assigning a new task to an AI, and if so, what are the additional benefits of this task and who does it benefit? If it is not a new task, then what process is it replacing, whose task does it replace, and who does it benefit? If you are replacing a human task, is it something for which computers have a comparative advantage? Does the AI free up your staff to take on more fulfilling human tasks? Does your new AI task augment your human staff's ability to succeed? Does your new AI task improve customer experience? Does your new AI task allow you to offer a better product? Does your new AI task expand your organization's capabilities?

Secondly, consider the objective(s) that the AI must optimize. Since an AI will ruthlessly optimize its objective, at the potential cost of other organizational objectives, list all competing objectives, and add them as constraints upon the AI as required. For example, you may give your AI an objective of increasing profit margin, but it may achieve that by making your product unaffordable to loyal customers and reducing your total revenue. So you could add constraints that the AI is not allowed to price more than 10% higher to current customers, and that the AI must not allow customer churn rates to increase.

There is more to this than merely considering the impacts upon your organization's internal business goals. Consider the negative externalities, the costs suffered by third parties as a result of the AI's actions. A thorough stakeholder analysis will identify the impact of the AI's task and objectives upon all stakeholders, both internal and external, not just the organization and its customers.

Pay particular attention to situations involving vulnerable groups, such as persons with disabilities, children, minorities, or to situations with asymmetries of power or information. List the stakeholders matching this definition and define the protections your organization will give to each group. Then based upon your stated protections, add constraints to your



AI that limit its ability to make decisions that negatively impact these groups. For example, you may add a constraint that your AI may not recommend adult-only products, such as alcohol or violent movies, to minors.

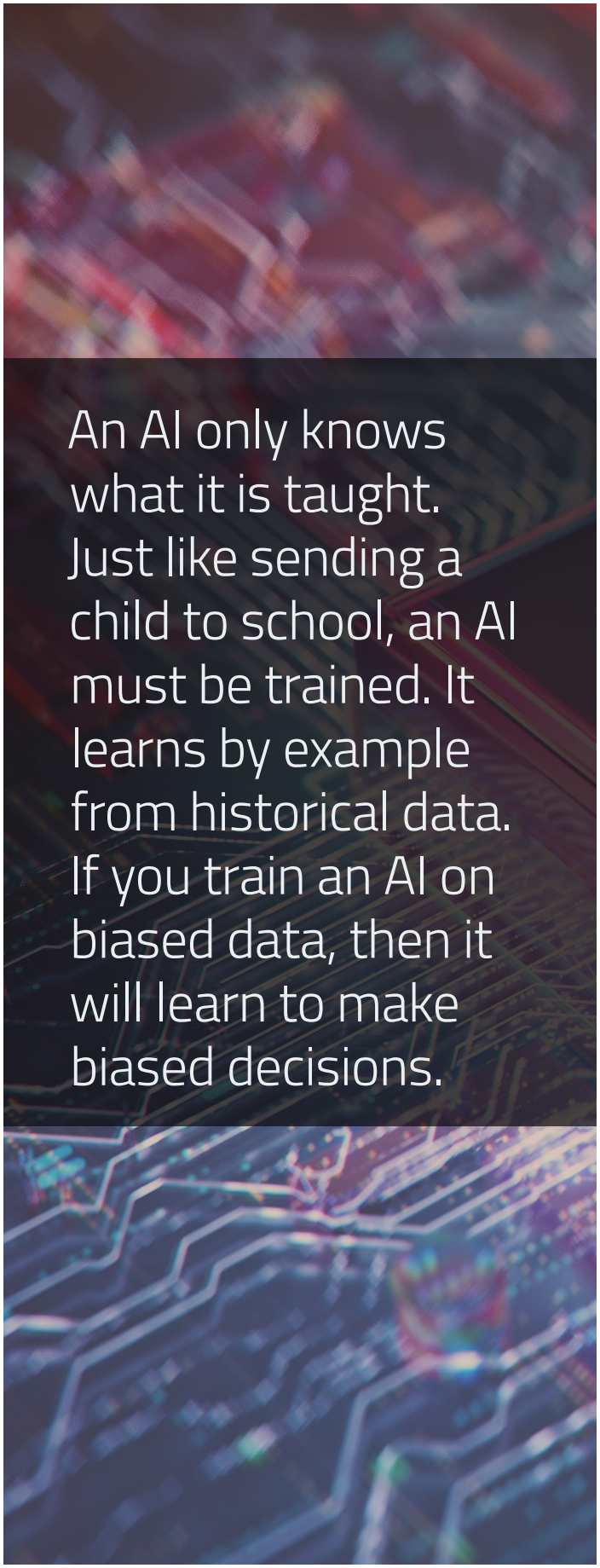
Consider the long-term impact of short-term objectives. Since your AI will ruthlessly optimize its ability to meet its stated objectives, it may make decisions that meet short-term objectives without regard to the long-term costs. This includes making a short-term profit at the long-term cost of reputation risk and brand value. It can also include actions that individually have low impact, but cumulatively, over extended time periods develop into problems. For example, an AI that is given the objective to write as many loans as possible, subject to no defaults in the next 12 months, may achieve this by issuing predatory loans to vulnerable borrowers who cannot afford the loans once introductory low-interest offers expire, with subsequent loan defaults after 18 months.

In summary, the general principle to apply is that your AI's actions should have a net good to society.

Principle 2: Fairness

The dictionary definition of "fairness" is, "impartial and just treatment or behavior without favoritism." Fairness is the opposite of unfair bias and unfair discrimination. In most cases, unfair bias involves one or both of the following conditions:

1. treating people differently (when you shouldn't!), and/or
2. entrenching historical disadvantage.



Most countries around the world have laws protecting against some forms of discrimination. The details will vary from country to country, from state to state within a country, and vary according to the context (e.g., employment discrimination protections for gender versus accepting the differences in healthcare needs by gender). Protected attributes around the world include gender, race, ethnicity, color, language, religion, disability, age, sexual orientation, pregnancy, political opinion, medical record, criminal record, marital or relationship status, trade union activity, and genetic information. It goes without saying that you should obey the law with regards to protected attributes.

While the illegal use of protected attributes can lead to expensive fines, reputation damage is not restricted to protected attributes. The inappropriate use of sensitive attributes can also damage your organization's reputation. Sensitive attributes include protected attributes but extend to attributes that your customers and society consider unfair or discriminatory. An attribute may be considered sensitive when it identifies a group or individual that is vulnerable or disadvantaged, where there is an asymmetry of power or information. This can include age, disability, language, socio-economic status, and many more. For example, some businesses use price elasticity modeling as part of setting prices. This often leads to them charging huge profit margins on elderly customers because older people do not typically shop around as much, nor churn as often. In many societies it is considered unethical to charge higher prices and profit margins to society's most vulnerable members. Many would consider it unethical to take advantage of a vulnerable grandmother. Some features could cause reputation risk if your use of them became public. For example, imagine the media headlines if you built an AI that used supply and demand data to decide to increase prices to $10 for a bottle of water after a devastating hurricane!

Ethical considerations go beyond the concerns of potential costs to one's self (such as fines or reputational damage) and extend to doing the right thing for its own sake, treating people impartially and fairly, simply because you believe it is the right thing to do.

Data scientists, the people who build and train AIs, have similar skills to teachers and lecturers. Just as we don't use textbooks from a hundred years ago, we shouldn't train AIs on data that contains unfair outcomes. If the historical data contains examples of poor outcomes for disadvantaged groups, then it will learn to replicate decisions that lead to those poor outcomes.

Data should reflect the diversity of the target population with which the AI will be interacting. Bias can also occur when a group is underrepresented in the historical data. If the AI isn't given enough examples of each type of person, then it can't be expected to learn what to do with each group. Decisions should not be the result of a metaphorical coin toss. Be cautious of small sample sizes. Since an AI relies only on historical data to learn how to make decisions, it will make much less reliable decisions for groups that are underrepresented in the training data, resulting in higher error rates and more cases of unnecessary negative decisions upon members of these groups.

When an AI learns to repeat human mistakes, it entrenches that behavior for the future. For example, if an AI is trained on data collected during a period in which women were less likely to be successful when applying for a job (possibly due to human bias or traditional gender roles that have become less common in recent times), then the AI will learn to prioritize men for recruitment.

> "The real safety question, if you want to call it that, is that if we give these systems biased data, they will be biased"
>
> **John Giannandrea, Google, "Forget Killer Robots—Bias Is the Real AI Danger", MIT Technology Review**

Bias can be direct or indirect. Direct bias is when sensitive attributes affect decisions and outcomes (e.g., when gender is used to determine who should be hired). Indirect bias is when another attribute is used as a proxy for a sensitive attribute, when that attribute is strongly correlated with the discriminatory attribute. For example, in the United States, a person's race is strongly correlated with their residential address. That is why U.S. banks are not allowed to use zip codes when making lending decisions. Indirect bias is a trap that some employers have fallen for, not realizing that the text in a resumé may contain words that only a female would use. But sometimes these correlated attributes may measure intrinsic differences that are truly related to outcomes. For example, it makes sense that the credit risk of a loan applicant is better the higher the applicant's income, yet income is often correlated with sensitive attributes such as gender, race, and disability. When considering indirect discrimination, you will need to use your common sense and judgment to separate true and fair effects from unfair bias.

Just because something happened in the past, doesn't mean that it will happen in the future or that you'll want it to continue to happen into the future.

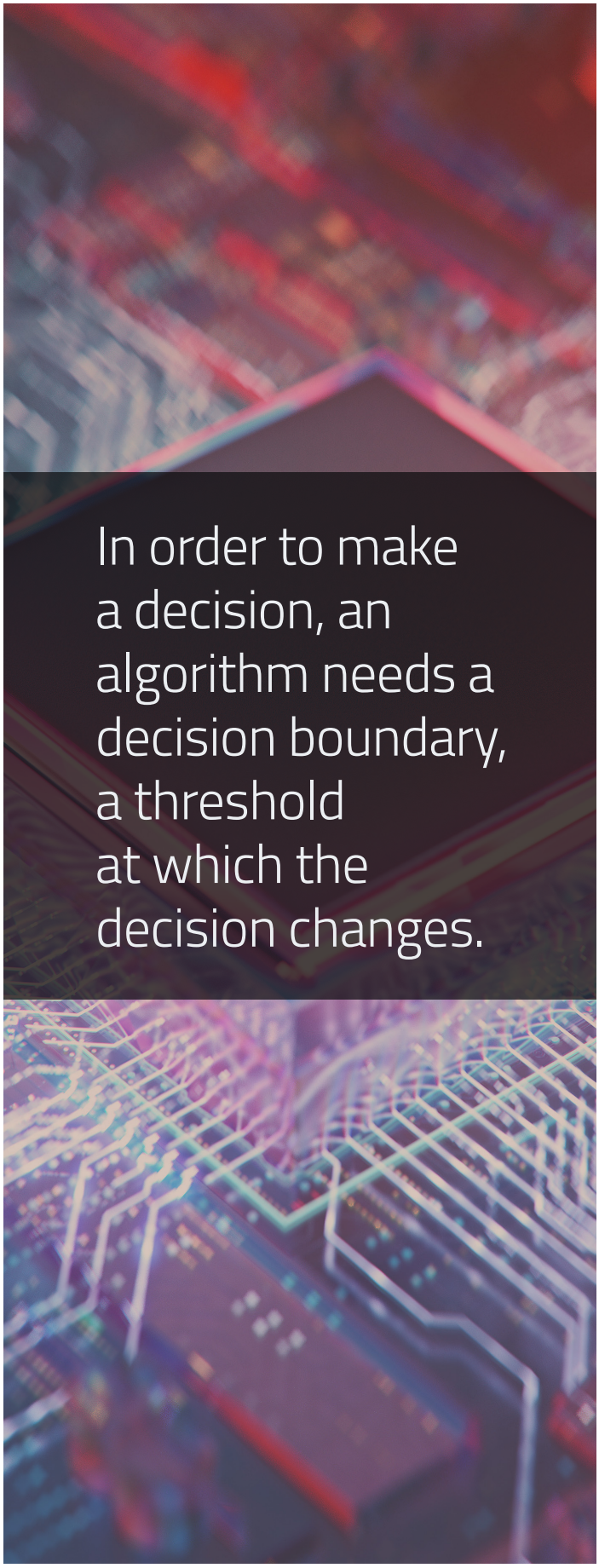
The good news is that while both humans and AIs can exhibit biased behavior, with AIs it is easier to detect and remove bias than with humans. Since an AI will behave the same way every time it sees the same data, you can run experiments and diagnostics to discover AI bias. There is no single generally accepted measure of algorithmic bias. Arvind Narayanan, an associate professor at Princeton lists 21 different definitions for detecting fairness in an algorithm. That's too many to list in this briefing, but here are several of the more popular measures used for detecting algorithmic bias for a protected group:

1. **Calibration:** whether the algorithmic predictions or decisions are accurate, including across groups.
2. **Statistical parity:** whether the algorithmic predictions or decisions are the same, on average, between groups.
3. **Equal opportunity:** The same outcomes, everything else being equal, no matter what the value of the sensitive characteristic. The sensitive feature has no effect on algorithmic outcomes (e.g., the probability of a male and a female getting a promotion is the same, so long as all other attributes are identical).
4. **Equal accuracy:** whether an algorithm is just as likely to be correct when classifying a person, regardless of their membership of a protected group (e.g., a product recommendation algorithm is just as likely to find the right product for a male as for a female).
5. **Equal false positive rate:** whether an algorithm is just as likely to make a mistake of identifying a person as a positive, regardless of their membership of a protected group (e.g., whether the probability of an unsuitable job candidate being selected for a job interview is the same regardless of their gender).
6. **Equal false negative rate:** whether an algorithm is just as likely to make a mistake of not identifying a person as a positive, regardless of their membership of a protected group (e.g., whether the probability of a prison inmate being incorrectly rejected for parole is the same regardless of their race).

You may think that the most ethical approach will be to ensure that your AI passes all of the measures of algorithmic fairness, but this isn't possible. Chouldachova's impossibility theorem implies that unless the model makes perfect predictions (0% and 100% probability), or the group base rates are equal, it is not possible to satisfy multiple fairness criteria at once. So, in the end, the decision is left in your hands - you must decide which type of fairness applies to AIs within your organization.

Another way to define discrimination, particularly in regulated situations, are two doctrines in anti-discrimination law:

1. Disparate treatment is one kind of unlawful discrimination in U.S. labor law. In the United States, it means unequal behavior toward someone because of a protected characteristic.
2. Disparate outcome occurs when policies, practices, rules or other systems that appear to be neutral result in a disproportionate impact on a protected group.



Just as with the various measures of algorithmic fairness, and depending upon the detail of how the requirements are worded, it can often be impossible to simultaneously meet both disparate treatment and disparate outcome requirements.

So far, we have discussed discrimination against groups, but discrimination and fairness can also be considered at the level of individual people. It is possible to have the same outcomes between groups but that this fairness of outcomes does not apply for each and every individual decision within that group. Some measures of algorithmic fairness can only be applied at a group level.

Issues may also arise due to personalization.

This threshold decides whether a loan is granted to an applicant, or which product is offered to a customer. It means that there will always be times when two almost identical people face different outcomes. Maybe the difference is that one earns just $10 per month more than the other, but that higher income crosses a boundary that makes them meet the acceptance criteria, while their near-doppelganger does not. Is it fair that two almost identical individuals face different outcomes? Maybe yes, or maybe no, but ultimately you cannot completely remove that possibility. What you can do is to try to set the decision boundaries objectively.

Finally, note that the word “discrimination” originally meant, “the recognition and understanding of the difference between one thing and another,” but in modern times we mostly hear about unfair discrimination (e.g., denying opportunities to people because of their race or gender). Some discrimination is good. For example, certain healthcare protocols vary due to age and gender, as what can be beneficial for one patient can be dangerous for another! You want your AI to discriminate in a positive manner, to understand and discriminate between constructive decisions versus destructive decisions.

In summary, the general principle to apply is that your AI’s actions should avoid discriminating on sensitive features, and avoid entrenching historical disadvantage.

Principle 3: Disclosure

One of the four fundamental principles of ethics is respect for autonomy, an obligation to respect the autonomy of other persons, to respect the decisions made by other people concerning their own lives. Sometimes this is called the principle of human dignity. It includes not only the duty not to interfere with the decisions of competent adults, but also a duty to empower others with whom we interact. Applying this to AI ethics, we have a duty to disclose to stakeholders about their interactions with an AI, so that they can make informed decisions.

> "AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills."

source: Independent High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI"

AI systems should not represent themselves as humans to users. Disclose to stakeholders when they are interacting with an AI versus when they are interacting with a human. Where practical, give the choice to opt out of interacting with an AI. For example, giving the option to speak to a human can be particularly helpful for customers when an unusual problem has arisen that the AI cannot fix, where the customer needs the empathy and creative problem solving that only another human can provide.

An AI system's capabilities and limitations should be communicated to AI stakeholders in a manner appropriate to the use case and appropriate to the stakeholder. This may include disclosing the accuracy of the AI or the edge cases where the AI's capabilities are limited. Disclose the implications of interacting with an AI and the rights and obligations of both parties. This will enable realistic expectation setting. Explain the degree to which an AI system influences the decision-making process and the rationale for deploying an AI.

Whenever an AI's decision has a significant impact on people's lives, it should be possible for those people to demand a suitable explanation of the AI system's decision-making process. The explanation should use human-friendly language, and the technical concepts communicated at a level tailored to the knowledge and expertise of the person. In some regulatory domains this is a legal requirement, such as the EU's General Data

Sufficient information should be provided that a stakeholder can make an informed decision about whether the decisions that affect them are reasonable, robust, and trustworthy. This information may include, but is not limited to:

- **Listing which data fields were used in the process**
- **The patterns and rules that the AI applied to the data values, and**
- **The specific data field values that were most important in making a decision**

Protection Regulation (GDPR) "right to explanation" and the "adverse action" disclosure requirements in the Fair Credit Reporting Act (FCRA) in the U.S. When practical, extend this right to include the ability to appeal an algorithmic decision. Disclose the availability and process of this ability to appeal.

In order for stakeholders to make informed choices, an AI's processes and decisions must be explainable. An AI must be able to provide human-friendly explanations of its decision-making process and the reasons for each decision it makes.

This does not necessarily imply that information about business models and intellectual property related to the AI system must always be made openly available. The disclosure of too much detail may enable and empower adversarial attacks upon the system.

In summary, the general principle to apply is that you should disclose sufficient information to an AI's stakeholders so that they can make informed decisions.

Principle 4: Governance

An organization's governance of AI is its responsibility to ensure that its AI systems are secure, reliable and robust, and that appropriate processes are in place to ensure responsibility and accountability for those AI systems.

Audit firm EY warns that your AI "can malfunction, be deliberately corrupted, and acquire (and codify) human biases in ways that may or may not be immediately obvious. These failures have profound ramifications for security, decision-making and credibility, and may lead to costly litigation, reputational damage, customer revolt, reduced profitability and regulatory scrutiny."

Since AIs are built by humans, and humans are fallible, AIs may contain human errors.

- Reduce the risk of human errors by automating the algorithm construction and training
- Standardize the process by which AIs are trained, ensuring that the process is reproducible
- Apply guardrails to the training of AIs to ensure best practices are followed and prevent errors

Like any other technology, AI can be used for ethical or unethical purposes, and AI can be secure or dangerous. With the possibility of negative outcomes from AI failures comes the responsibility to manage AIs, apply high standards of governance and risk management.

An organization should have clearly stated ethical guidelines that apply to all its AI projects. A stakeholder analysis and impact assessment should be carried out prior to the development of a new AI project. Risk evaluation should identify material risks, where an AI's decisions or an AI failure could have a significant impact upon a stakeholder. In situations where such risks exist, evaluate whether those risks can be reduced or are justified within the context of the ethical guidelines.

Humans must be responsible for, and have accountability for, the AIs they design and deploy. Human oversight helps to ensure that an AI system is performing as expected without unforeseen adverse effects. The comparative advantage of humans over computers in general knowledge, common sense, context and ethical values means that the combination of humans plus AIs will deliver better results than AIs on their own. When the risks are high, such as with life or death medical decisions, adequate governance may require humans to be involved in each decision-making step, with power to intervene and override. In many cases, though, this is neither practical nor desirable. For example, it is not practical to have a human sign off each decision in a process that must scale to millions of customers. In such cases, the role of humans is to manage the design of the AI, monitor its operation, and set policies and business rules for when and how to use the AI in different situations.

Facilitate the traceability and auditability of AI systems. Document the data sets and the processes that trained the AI and the pipeline within which the AI is deployed. In order to maintain a high standard of documentation, automate the documentation to ensure consistency and reduce human error. Keep an audit trail of the checks that were applied to ensure that the AI performed consistently with business requirements and ethical guidelines, and that the system is robust to errors. This also applies to the decisions that an AI makes. Record each AI decision, plus its reasons, to enable the identification of reasons why an AI made an error and to prevent a repeat of those errors in the future.

As with all software, AIs should be protected against vulnerabilities that can be exploited by hackers. Attacks may target the source data, the model design, or the infrastructure on which the AI is deployed. For AIs to be secure, identify vectors of potential malicious attacks and take steps to prevent and mitigate these risks. AI systems must protect the privacy of people whose data they use.

In summary, the general principle to apply is that where there is risk, apply high standards of governance over the design, training, deployment, and operation of AIs.

Conclusion

The application of AI ethics can improve your organization's effectiveness, reducing regulatory risk, reputation risk and providing a net benefit to society.

"Firms with strong positive reputations attract better people. They are perceived as providing more value, which often allows them to charge a premium. Their customers are more loyal and buy broader ranges of products and services. Because the market believes that such companies will deliver sustained earnings and future growth, they have higher price-earnings multiples and market values and lower costs of capital."

Robert G. Eccles, Scott C. Newquist, Roland Schatz, "Reputation and Its Risks", Harvard Business Review, February 2007

Your first step to ethical AI is to develop an AI Ethics Statement that will apply to all AI projects and deployed AIs across your organization. That policy will clearly state detailed guidelines for how the principles of ethical purpose, fairness, disclosure, and governance are to be applied and will reflect the values of both your organization and society. DataRobot offers a workshop that introduces your executives to the principles of AI ethics, then assists them to write their first AI Ethics Statement.

DataRobot's automated machine learning platform can be a valuable tool to implement ethical AIs. With guardrails, standardization, reproducible model training, deployment, and human-friendly explanations, DataRobot is the industry leader, listed as a Visionary in the 2019 Gartner Magic Quadrant for Data Science and Machine Learning Platforms. Our AI Success teams work with your executives and data scientists to provide advice and training so that you are in control of the values and governance of your AIs.

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.

Learn more at **datarobot.com**

v05152019.0709