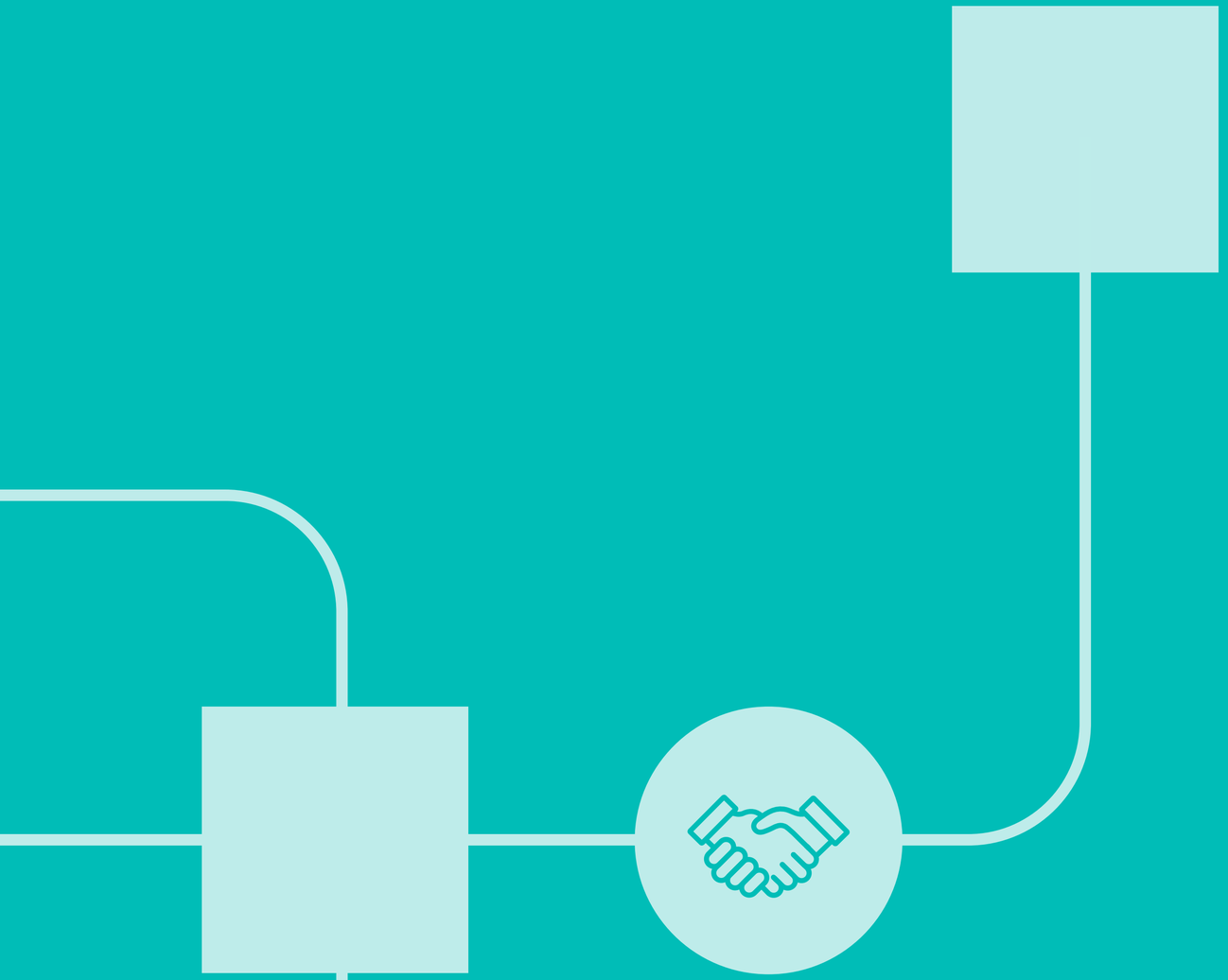


data
iku

AVOID AI PROJECT FAILURE:

How to Build Trustworthy AI Systems

According to the National Institute of Standards and Technology (NIST) “Trust and Artificial Intelligence” report,

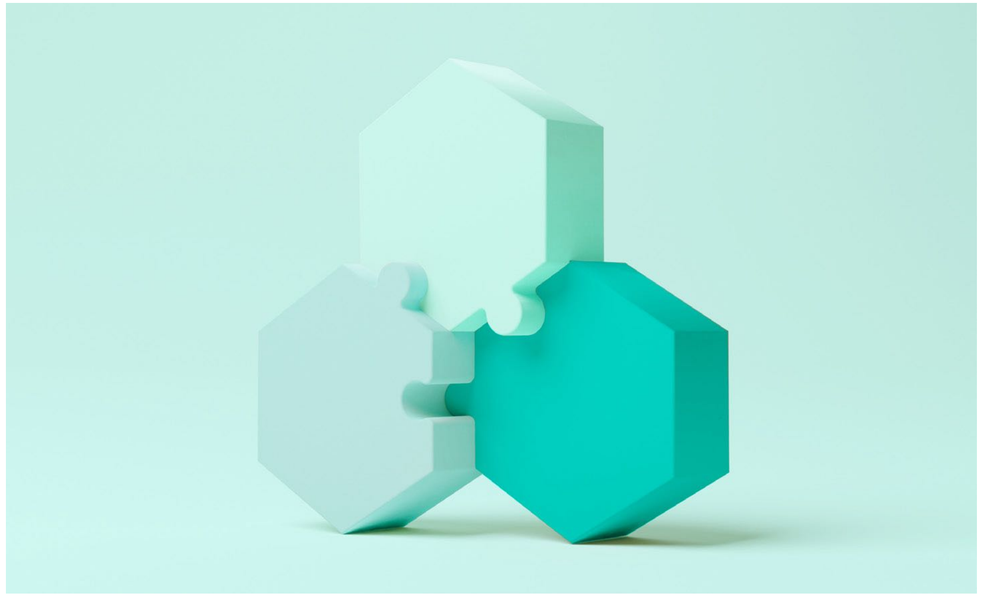
> “AI has the ability to alter its own programming in ways that even those who build AI systems can’t always predict.”[1]

Unlike spreadsheets, dashboards, and ERP systems, AI software can automatically and autonomously change its coding via retraining, adaptive learning, or reinforcement learning. AI also operates thousands of times faster than human workers, so when AI software does harm, it can do it at a scale that most managers have never seen before.

For these reasons, AI practitioners and developers are creating new processes to make AI trustworthy. But trustworthiness is very relative, which is what makes developing this type of technology very challenging. Trustworthy AI is important when it comes to driving frontline user adoption and managing risk (both of which are essential elements to avoiding analytics and AI project failure).

This flipbook will dig into both of those reasons, as well as provide an overview on:

- **Types of risk, harm, and bias**
- **Key attributes of trustworthiness** (i.e., accuracy, reliability, explainability)
- **Ways to put trustworthy AI into practice**

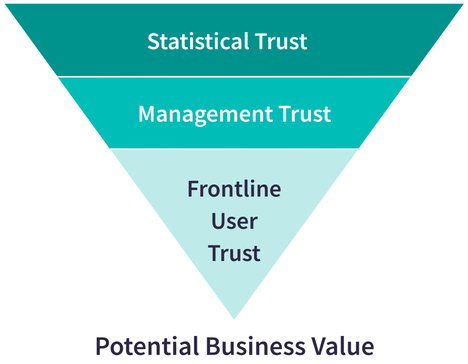


1 *https://www.nist.gov/publications/trust-and-artificial-intelligence*

Trustworthy AI for Frontline User Adoption

Even the most accurate AI model can't generate business value if it's not used.

What AI developers consider trustworthy differs from what management considers trustworthy, and both are often different from end user trustworthiness. Two of these three levels are non-technical, subjective decisions made by people.



> *'Trustworthy' isn't an attribute of data, an AI model, or an AI app. It's a relationship between a person and data, a model, or an app created solely by the person."*

The history of AI is full of statistically good models that were business failures because frontline users didn't trust them.

A Real-World Example: Duke University[2]

A Duke University hospital admissions team spent a year developing an AI application to help emergency room staff decide if a patient with chest pain should be admitted to the intensive care unit (ICU). Increasing the accuracy of those decisions would decrease ICU workload, improve patient care, and reduce costs for both the hospital and patients.

However, doctors objected to being told how to do their job, and busy emergency room staff rejected the extra steps in the admission process required to use the application. It failed in just three weeks.

2 *https://sloanreview.mit.edu/article/ai-on-the-front-lines/*

Trustworthy AI for Risk Management

While user adoption promotes the potential upside of AI, risk management focuses on the potential downside. A first step here is defining “downside” — based on your organization's principles and goals — in a way that it can be measured.

For some organizations, it might be financial, such as revenue or cost. For others, it may include brand reputation or social fairness. Every organization and context might be different and they can change over time, such as emphasizing costs during economic contractions and revenue during growth periods.

A Real-World Example: Zillow

Zillow’s residential home price model is a cautionary tale of how risk changes with time and context.[3] Initially, its predictions were displayed on Zillow.com so visitors could see what their home (and others) might be worth.

Over time, the model improved and a large team of internal property experts began using it to buy and flip homes for a quick profit. That worked so well that management removed the human experts from the decision-making process to speed things up, bought 3,000 homes without sufficient human oversight, lost $570 million, and laid off 2,000 people.

---

3 *https://www.bloomberg.com/news/articles/2021-11-08/zillow-z-home-flipping-experiment-doomed-by-tech-algorithms*

Types of AI Risk to Consider

One may define risk as:[4]

Risk = potential harm * likelihood to occur

Most industries, organizations, and departments have different thresholds for each component. The Hippocratic Oath's "first, do no harm" in medicine emphasizes harm regardless of its magnitude. The U.S. Federal Trade Commission (FTC) defines risk the same way for some types of harm, such as to civil liberties, where no amount of harm is allowed. In most practical situations, risk is weighed against benefits:

Unfair = potential harm * likelihood to occur – benefit

Three hundred people in the U.S. drown in bathtubs each year and yet we still have bathtubs. Over 80 people die in car accidents per day, yet 17 million new cars are sold each year.

This is both because the likelihood is small and the benefits are big. The FTC uses this for most AI risks, equally weighting benefit and harm: "To put it in the simplest terms… [an AI] practice is unfair if it causes more harm than good."[5]

---

4 *https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf*
5 *https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai*

Types of Harm to Consider

When evaluating trustworthiness of your AI systems, determining which types of harm are most critical depends on your organization's principles and goals. Types of harm include:[6,7]

INDIVIDUALS

- Physical harm or death
- Financial
- Reputation
- Privacy
- Civil liberties
- Civil rights
- Disparate treatment of individuals of sensitive or protected classes
- Disparate outcomes for sensitive or protected classes

ORGANIZATIONS

- Financial
- Compliance
- Legal
- Reputation

SOCIETY

- National security
- Public markets
- Financial & credit system
- Supply chain (e.g, food, medicine)
- Voting, education, energy, water, transportation, healthcare, communications, and navigation (GPS) infrastructure

---

6 *https://www.ftc.gov/reports/combatting-online-harms-through-innovation*
7 *https://www.mckinsey.com/business-functions/quantumblack/our-insights/confronting-the-risks-of-artificial-intelligence*

Today, many AI practitioners focus on only a few potential harms like an organization's finance and compliance. Developers, managers, and users will need to consider more types as AI becomes more widespread, and make explicit design-time decisions on which risks are important and how to measure them in production. If they're not defined in a measurable way, then they're just documentation — and that's insufficient for generating trust.

Harm from AI systems is typically either never considered by developers or an unintended consequence. A common way that it occurs is from bad data.[8] Three ways that AI models can generate harm are by using biased datasets:

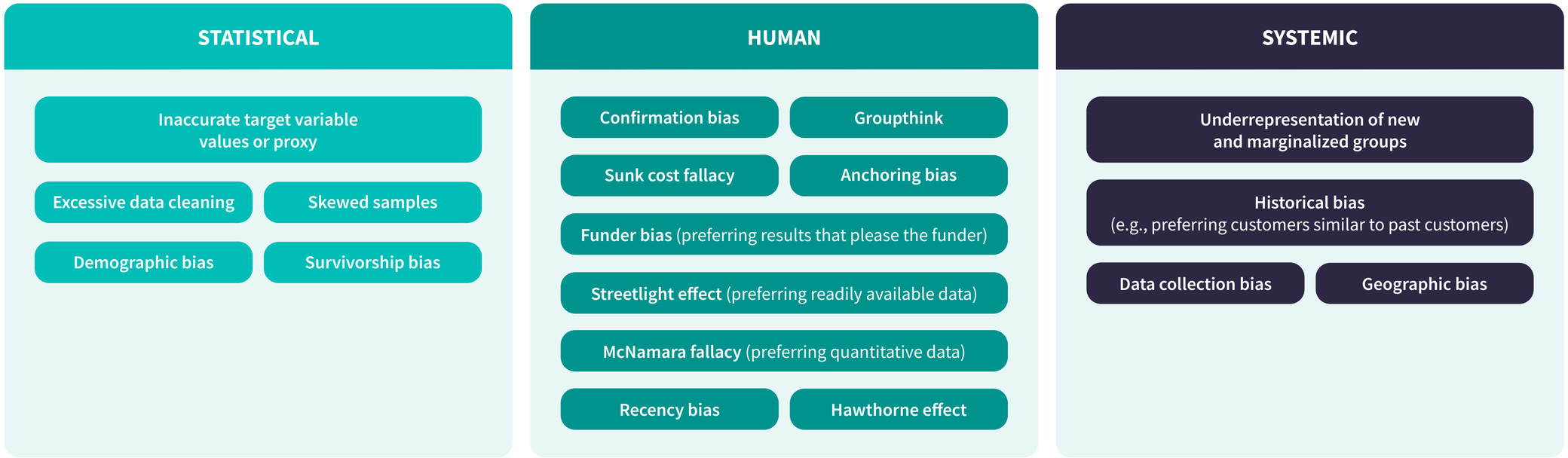
- Feature and target data that **does not accurately represent the truth**
- Inference (also known as prediction or scoring) datasets are **significantly different than training data**
- **Malicious internal or external actors** introduce bias in datasets

The next section reviews types of bias to guard against.

8 *https://storage.googleapis.com/pub-tools-public-publication-data/pdf/0d556e45afc54afeb2eb6b51a9bc1827b9961ff4.pdf*

Types of Bias to Consider

The question isn't whether your training data, model, or inference data is biased (it is) — it's how and whether it's important to you. Three high-level sources of bias in AI systems are:[9,10]



9 *https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf*
10 *https://www.dwt.com/-/media/files/blogs/artificial-intelligence-law-advisor/2022/03/nist-sp-1270--identifying-and-managing-bias-in-ai.pdf*

A Real-World Example: Statistical, Human, and Systemic Bias in Healthcare

A widely used commercial model estimates how sick a healthcare patient is. Scarce resources such as teams of dedicated nurses and extra primary care appointment slots are then allocated based on that.

The model, however, used next year's healthcare costs as a proxy for sickness magnitude. The developers introduced streetlight bias by using readily available billing data. The problem is that Black patients have less access to medical insurance than white patients and thus are billed less for the same degree of sickness.

Thus, for the same level of sickness, the model inferred that white patients are sicker and they get additional care. Billing is a bad target variable proxy due to systemic inequalities in the healthcare marketplace. Researchers estimate that if this bias were removed, the number of Black patients getting additional care would increase by 160% or more than double. [11]

Note that not all biases are harmful, even for protected classes. Many retail apparel product recommendation models, for example, are more accurate for women than for men, but that's not considered unfair. Each organization needs to decide which biases, harms, risks, and sensitive subgroups are important to them.

11 *https://www.ftc.gov/system/files/documents/public_events/1548288/privacycon-2020-ziad_obermeyer.pdf*

Metrics for Detecting Bias in Sensitive Subgroups

Dataiku — the platform for Everyday AI — provides explainability features that help analytics and AI project builders (and their stakeholders) stay aligned to their organizational values, increase trust, and eliminate bias.

There are many metrics for detecting bias in sensitive subgroups in training, inference, and prediction data including:[12,13]

1. **Demographic representation:** Does a dataset have the same distribution of sensitive subgroups as the target population? The Kolmogorov-Smirnov test could be used.
2. **Demographic parity:** Are model prediction averages about the same overall and for sensitive subgroups? For example, if we're predicting the likelihood to pay a phone bill on time, does it predict about the same pay rate for men and women? A t-test, Wilcoxon test, or bootstrap test could be used.
3. **Equalized odds:** For boolean classifiers that predict true or false, are the true positive and false positive rates about the same for sensitive subgroups? For example, is it more accurate for young adults than for the elderly?
4. **Equality of opportunity:** Like equalized odds, but only checks the true positive rate.
5. **Average odds difference:** The difference between the false positive and true positive
6. **Odds ratio:** Positive outcome rate divided by the negative outcome rate. For example, (likelihood that men pay their bill on time) / (likelihood that men don't pay their bill on time) compared to that for women.
7. **Disparate impact:** Ratio of the favorable prediction rate for a sensitive subgroup to that of the overall population.
8. **Predictive rate parity:** Is model accuracy about the same for different sensitive subgroups? Accuracy can be measured such as precision, F-score, AUC, mean squared error, etc.

Gini, Theil, and Atkinson indices have also been used to measure disparity. Which metric or test to use depends on the point in the AI product lifecycle and understandability by those consuming the information. Data scientists, for example, may be comfortable with a Theil index, but it may be too complex for business stakeholders and thus lower trust.

---

12 https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf
13 https://blog.dataiku.com/navigating-targeted-ad-bias-with-responsible-ai

Key Attributes of Trustworthiness

A wide variety of trustworthy AI attributes have been proposed but there is a consensus — from central banks to global system integrators to Microsoft and Google — that they include accuracy, security, explainability, and accountability.

Accuracy and Reliability

Accuracy applies to both data and models and, within data, to both features and target variables. For features, it might be monitored by the null value rate or changes in distributions. Target variable accuracy is always a concern, especially when it's generated manually by people. People, like models, are biased.

For example, when labeling images (is it a cat, dog, car, etc.) people have a minimum error rate of 5%, and experienced radiologists contradict themselves about 20% of the time.[14]

Reliability refers to a model's accuracy being consistent under expected conditions for an expected amount of time. For example, is it expected to be reliable for a couple of days, a month, a year? Being transparent with stakeholders on the conditions and timeframe for reliability builds trust.

---

14 *Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018) Prediction Machines, Harvard Business Review Press*

Safety, Security, Resiliency

High-risk AI systems should be secured against malicious attacks from internal or external bad actors, such as changing target variable data with the intent to harm. They should also continuously monitor for such attacks and detect them as soon as possible.

The level of security and monitoring that's appropriate depends, of course, on the risk. An AI system cross-selling toothpaste is certainly lower risk than one detecting credit card fraud and thus may have different security

A Real-World Example: A Model to Identify Good New Customers

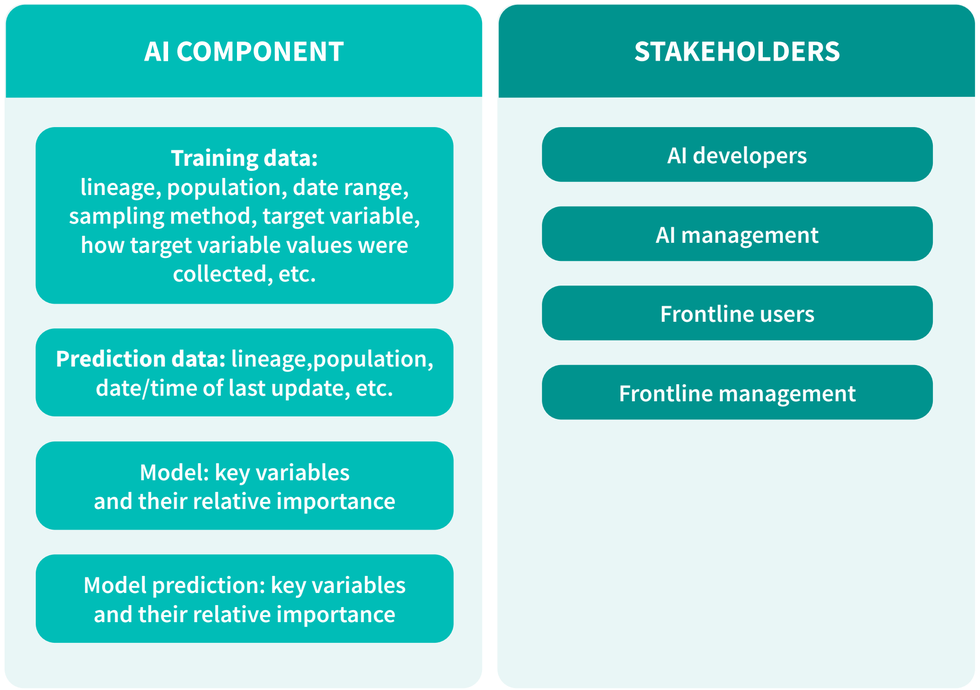
Resilience applies to sudden changes in the environment or usage. One of the variables a company looking to identify good new customers used was how many cars a person owned. The more cars they had, the more likely they were to be a good customer. That works in New York City since most people with multiple cars are wealthy.

The company expanded into the Southeast and applied the same model to people there. Model accuracy dropped quickly because in the rural Southeast low-income people tend to have many cars.

The company lost a lot of money and postponed their expansion plans. The meaning of a key variable — multiple cars — had changed and they did not detect it in time. If harm is done, bias is detected, or accuracy drops below acceptable thresholds, then trustworthy AI systems quickly detect it, alert stakeholders, shutdown, repair, or retrain.

Explainability

Explainability applies to both data and models, and — like trust — is in the eye of the beholder. There are four AI system components worth explaining and four kinds of stakeholders to explain them to:



Of course, data practitioners do not need all four AI system components for every AI app since risk can vary greatly. How explanations are presented can make a big difference in driving trust also, even for expert end users.

A Real-World Example: Harvard University Medical School Study

The study involved recommending radiation or surgery for lung cancer patients. When doctors were told that surgery had a 90% survival rate the first month, 84% of them picked surgery. When told that surgery had a 10% mortality rate the first month, only 50% selected surgery. Both explanations were accurate, and surgery was the right choice, but the results were very different.[15]

While the example above isn't AI per se, it neatly illustrates the relationship between how explanations are presented and trust. An explanation that's usable by one type of stakeholder may be inappropriate for another type, so some AI components may require multiple explanations.

15 *Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018) Prediction Machines, Harvard Business Review Press*

Accountability

Who is held responsible when harm occurs? Users and other stakeholders have an expectation that someone will be and that they're at an appropriate managerial level, not a developer. Clearly stating who is responsible in model and app documentation helps build trust.

Now for the Practical Applications

In this section, we get more concrete about the people, processes, and technology used to develop trustworthy AI.

People and Processes

Since trust drives adoption and adoption drives ROI, it shouldn't be a surprise that the same key driver of AI ROI is also a key driver of AI trust, namely, a multi-stakeholder approach with interdisciplinary development teams that gets frontline users involved early.[16,17,18]

Identify your frontline users, and not just a persona but specific people. Get a few of them involved in co-developing and reviewing your data, AI model, and AI app plans. Specifically, co-design A/B tests with them to prove the value and measure the risk of your models. Be careful to guard against funder bias by depending too much on the opinion of frontline user management rather than actual frontline users.

---

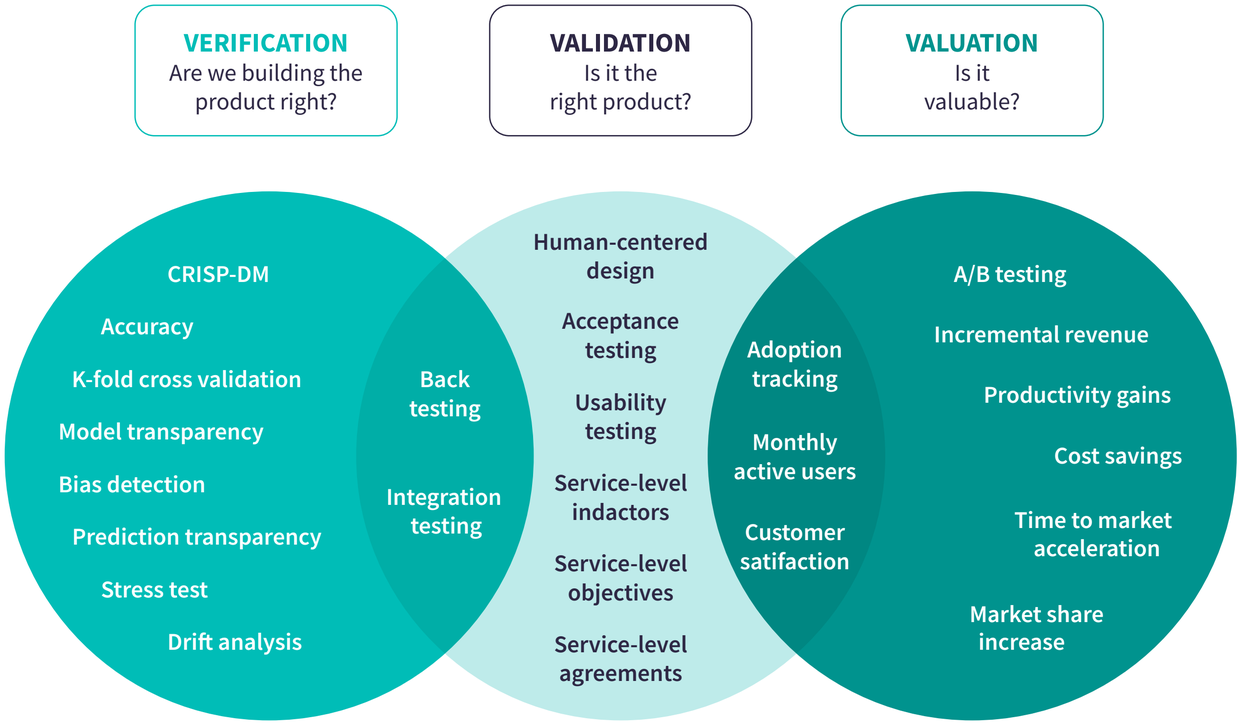
16 *https://sloanreview.mit.edu/article/ai-on-the-front-lines/*
17 *https://www.dwt.com/-/media/files/blogs/artificial-intelligence-law-advisor/2022/03/nist-sp-1270--identifying-and-managing-bias-in-ai.pdf*
18 *https://hbr.org/2019/07/building-the-ai-powered-organization*

Verification, Validation, and Valuation

Consider three levels of quality for data and AI products:

- **Data:** Accuracy and transparency
- **Users:** Usability and adoption
- **Business:** Value and ROI

Traditional software engineering maps the first two quality levels to verification and validation. We can add a third “V,” valuation. A few of the techniques in each are:



Audits

Like code reviews, even the best AI teams can benefit from audits. Three types of audits are:

- **Internal:** Performed by a separate team within the same company or organization. Facebook, who is at the cutting edge of AI technology, uses this approach extensively.[19]
- **External:** Hire an outside team to review data and AI products. Some conglomerates and holding companies use teams from another company with the same parent rather than external consultants to minimize legal issues and promote knowledge sharing.
- **Regulatory:** Audits performed by law enforcement agencies.

In order for audits to effectively build trust, there needs to be a degree of equality between AI developers and auditors. Auditors should have the same skill level and incentives as developers (perhaps even quotas), and not be influenced by funder or incumbent bias (i.e., regulatory capture).

Audits, like other risk management, should be ongoing and proactive. Waiting for users to detect problems erodes trust. There's a saying in software engineering that it's okay to have bugs, but it's not okay for the users to be the first to know about them. Audits may be triggered by changes in model accuracy, detection of bias, or the time since the last audit.

19 *https://www.wired.com/story/facebooks-red-team-hacks-ai-programs/*

Reporting

Audits, accuracy monitoring, bias detection, service-level objective compliance, and checking for discriminatory outcomes build trust when they are proactive and continuous.[20,21] Ideally, users should be able to review a history of such checks, the issues raised, harm detected, and their resolutions. Updating that documentation continuously and making it open to all potential stakeholders and users builds trust.

Most stakeholders don't need to know the internal details of an AI system. Increasingly, though, some users are other developers who do need details in order to trust data and AI products. Those details include:[22,23,24,25]

---

20 *https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai*
21 *https://committee.iso.org/ISO_37000_Governance*
22 *https://arxiv.org/pdf/1810.03993.pdf*
23 *https://arxiv.org/abs/1803.09010*
24 *https://martinfowler.com/articles/bitemporal-history.html*
25 *https://partnershiponai.org/about-ml-process-guide/*

Descriptive statistics are most useful when they're up to date, which is easy to provide in today's platforms via dashboards, statistical workbooks, and computational notebooks.

An extensive list of data and AI product characteristics to consider can be found in ISO 25012,[26] W3C data quality vocabulary,[27] and W3C data catalog vocabulary.[28]

---

26 *https://iso25000.com/index.php/en/iso-25000-standards/iso-25012*
27 *https://www.w3.org/TR/vocab-dqv/*
28 *https://www.w3.org/TR/vocab-dcat/*

Technology

Technology plays a key role in developing data and AI product trust. When done well, it serves as an objective, always-on watchdog that simplifies AI developers' daily work by eliminating tedious, manual tasks. Automated tests are more likely to be executed and reduce the likelihood that developers will find workarounds. In order to build AI trust, Dataiku provides a centralized place for the checks and balances outlined below:

1 Variable importance in
- **Models**
- **Individual predictions**
- **Nearest neighbor predictions**
- **Sensitive subgroups**

2 **Model stress tests**[29]

3 **Rapid deployment of new data and models**, shutdown of bad ones, and rollback to previous versions. Risk, for example, is higher for a model that takes three days to shutdown than for one that takes three minutes.

4 **A/B testing of models in production**

5 **Detecting data and model accuracy drift**

6 **Detecting data and model accuracy drift in sensitive subgroups** (i.e., bias)

7 **Ongoing backtesting**

8 **Triggers to remove a model from product**, rollback to a previous version, or automatically retrain

9 **Alerts** sent via email, SMS, Slack, Microsoft Teams, JIRA, Datadog, etc.

10 **Residual analysis:** Are inaccurate predictions random or is there bias?

11 **User adoption analysis:** Are the potential users who decide not to use the product random or is there bias? For example, is the adoption rate in rural hospitals much lower than that in suburbs?

12 **Prediction override analysis:** An "override" is when a user decides not to use a prediction and instead goes with their own judgment. Are overrides random or is there bias? For example, are overrides more common for expensive products, high-end customers, or life-threatening diseases?[30]

13 **Automatic dataset and model documentation generation**

14 **Automatic update** of dashboards, statistical workbooks, and computational notebooks

15 **Enterprise-wide data and model catalogs for assessing systemic risk**

There are a variety of variable importance methods including Shapley values, Individual Conditional Expectation (ICE), Local Interpretable Model-Agnostic Explanations (LIME), and BayesLIME which adds credible/confidence intervals to LIME. There is no one "right" method. It depends on what your stakeholders trust.

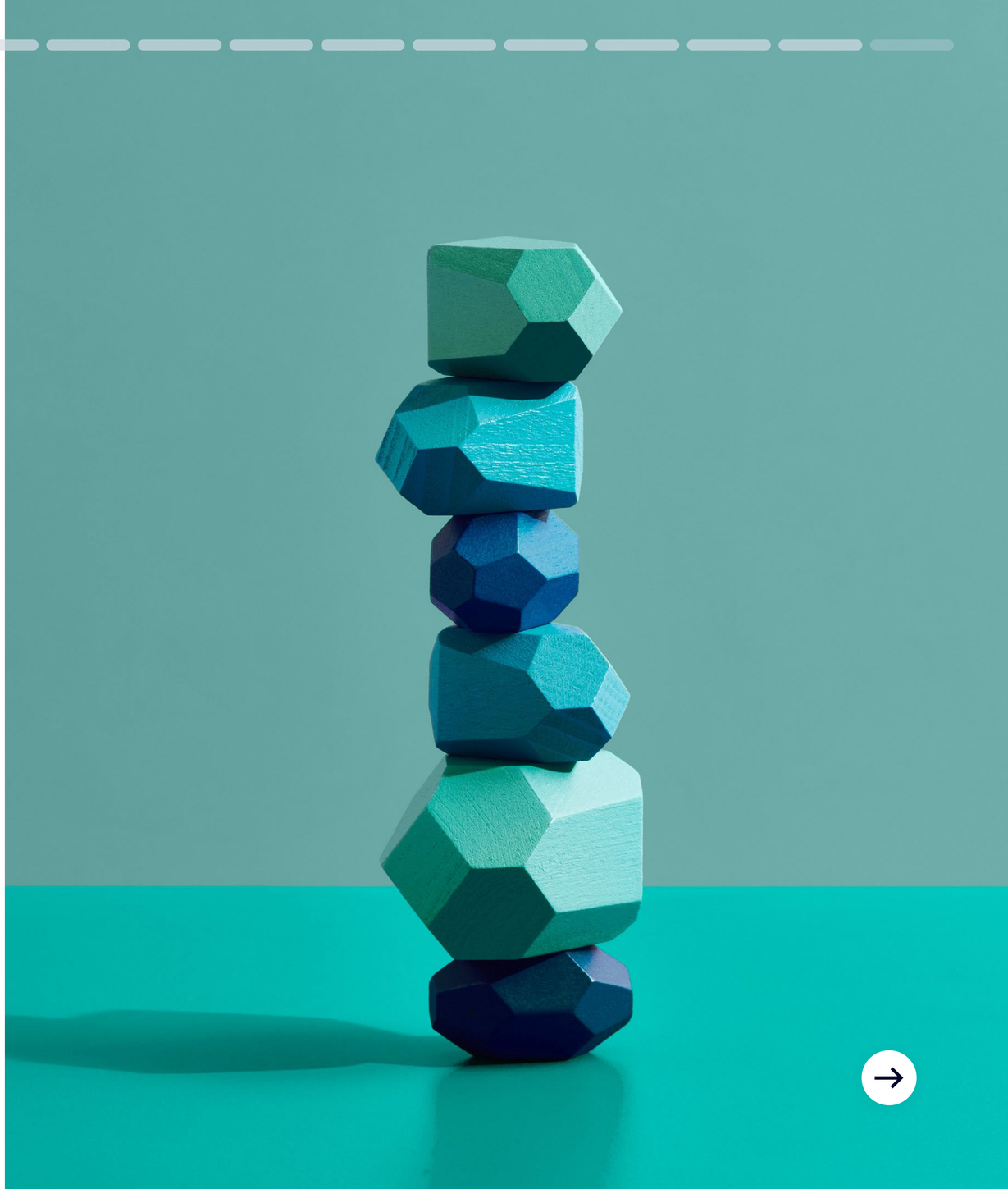
---

29 *https://blog.dataiku.com/the-hidden-force-stopping-your-ai-project-trust*
30 *Davenport, Thomas and Steven Miller (2022) Working with AI: Real Stories of Human-Machine Collaboration, The MIT Press*

Conclusion

Trust is a relationship between an AI product and a potential user, not a static attribute of the product. The user decides. An AI product doesn't need to be trusted by everyone to be successful. Grocery store recommendations, for example, are used by about 2% of customers but still generate big value.

Empathy, transparency, explainability and diligent monitoring build trust. As AI becomes more industrialized, practitioners can spend less time on data wrangling and bias-variance trade offs, and more time on understanding, quantifying, and measuring the harm and biases that users care about. A more human-centered approach will drive adoption, trust, and ROI.



Go Further on Trust and Explainability With Dataiku

Practice Responsible AI by understanding pipelines and interpreting model outputs to increase trust and eliminate bias.

LET'S GO