

Trustworthy and responsible AI in health care and life sciences

A leadership perspective

Mark Lambrecht, PhD, Global Head of Health and Life Sciences Advisory



Responsible and trustworthy AI involves the ethical, transparent, and fair application of AI technologies, ensuring they benefit all patients while minimizing risks and biases

National Academy of Medicine, USA - <https://nam.edu/artificial-intelligence-in-health-health-care-and-biomedical-science-an-ai-code-of-conduct-principles-and-commitments-discussion-draft/>

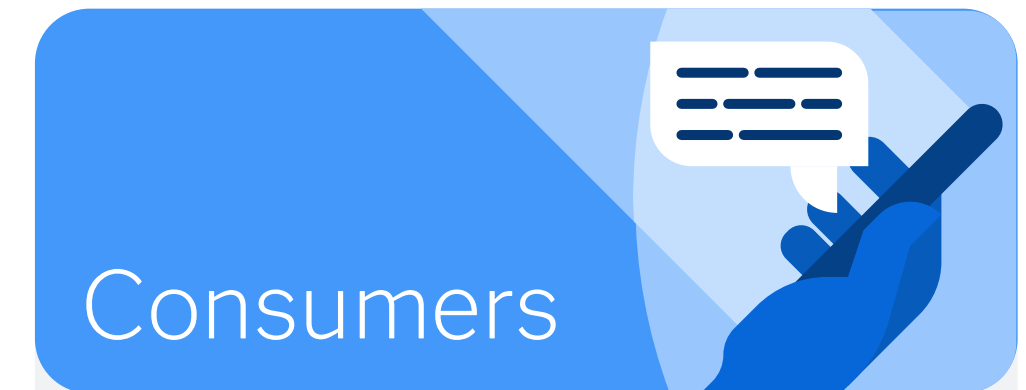
AI Impacts Everyone



- Can we make sense of our data?
- Do customers trust us?
- Are we complying with the law?

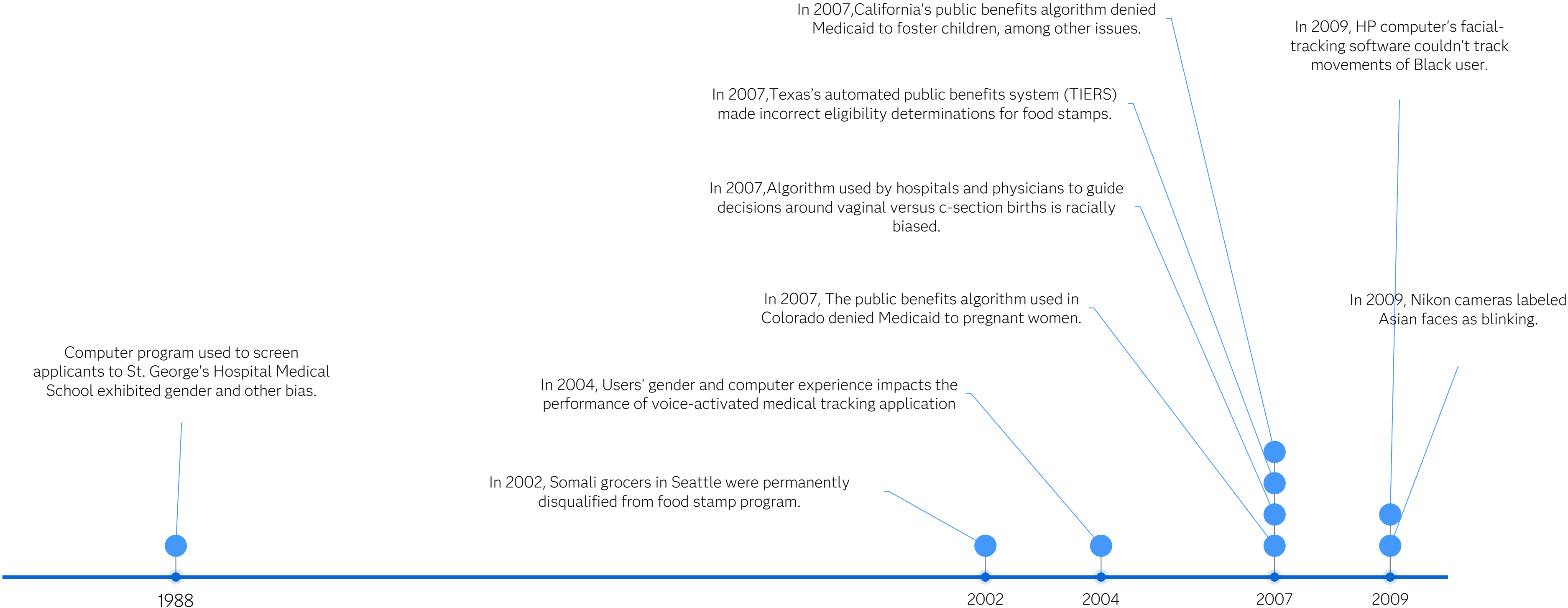


- Does AI make fair decisions?
- Is it transparent?
- Is it safe for consumers?

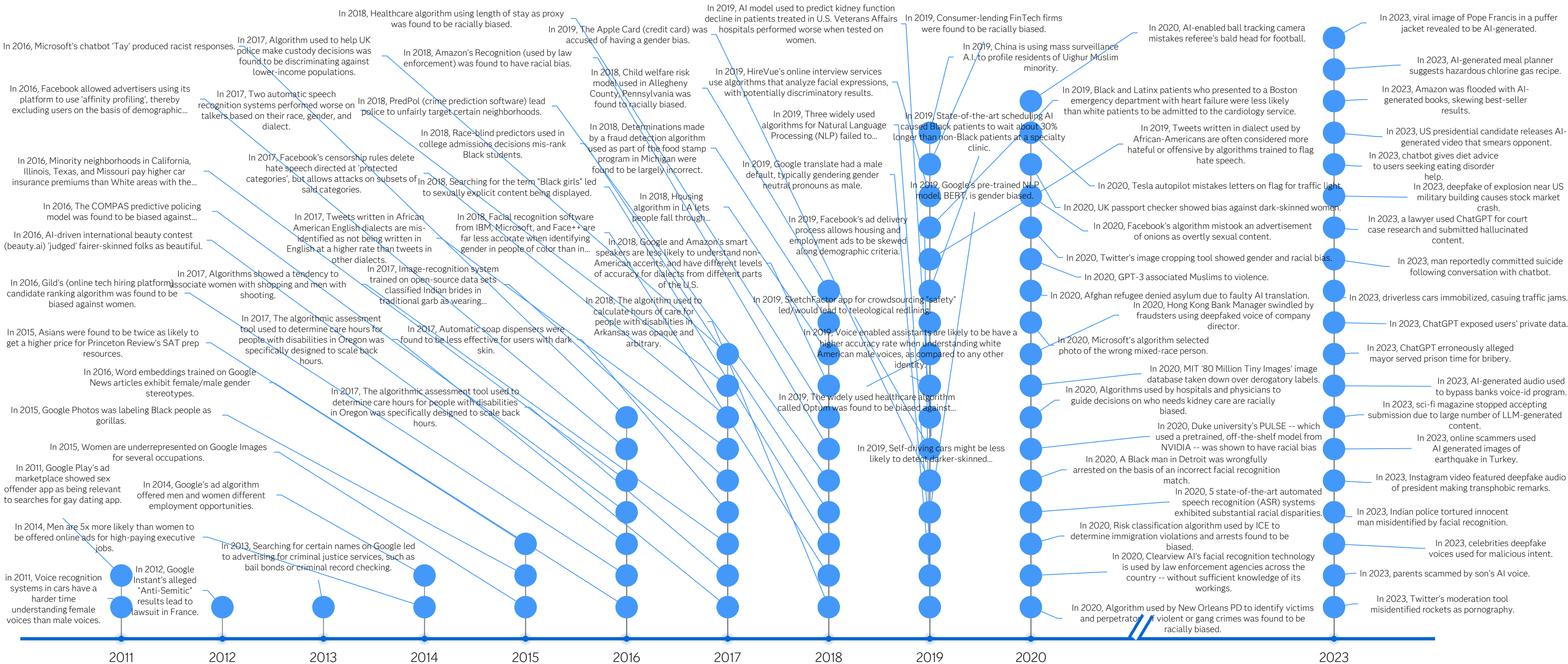


- Is my privacy protected?
- Am I being treated fairly?
- What can I change for a more positive outcome?

Unintended harms occur...



...now at MASSIVE scale



The Business Imperative

**A.I. is getting more powerful, faster, and cheaper—
and that's starting to freak executives out**

BY JEREMY KAHN
March 9, 2021 at 11:58 AM EST

**ChatGPT is biased and offensive,
creators admit**

OpenAI compares fine-tuning to training a dog

**IBM Abandons Facial Recognition Products,
Condemns Racially Biased Surveillance**

June 9, 2020 · 8:04 PM ET

**'There is no standard':
investigation finds AI
algorithms objectify
women's bodies**

Guardian exclusive: AI tools rate photos of women as more sexually suggestive than those of men, especially if nipples, pregnant bellies or exercise is involved

**Amazon built an AI tool to hire people but had to shut it down
because it was discriminating against women**

Isobel Asher Hamilton Oct 10, 2018, 5:47 AM

HEALTH / SCIENCE / ARTIFICIAL INTELLIGENCE

Hospitals use a transcription tool powered by a hallucination-prone OpenAI model

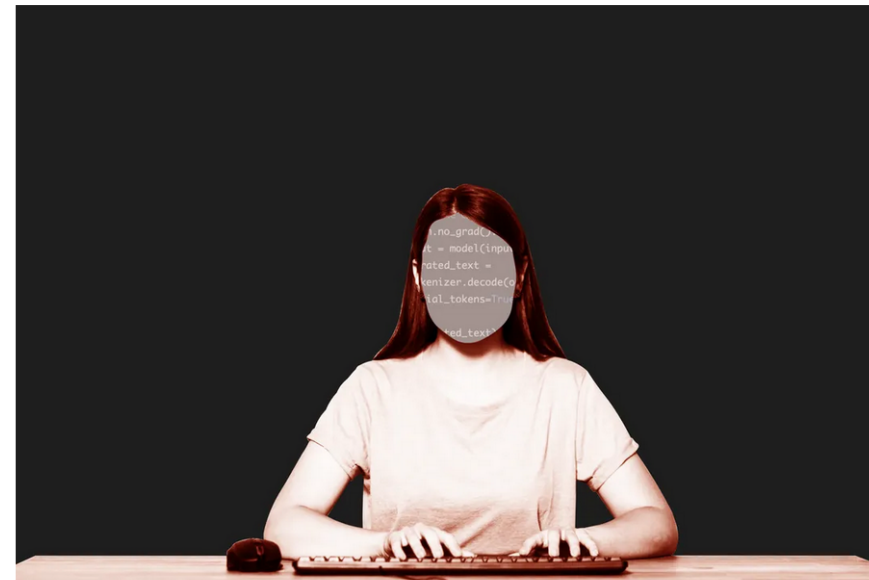


Image: The Verge

/ Researchers have found that Whisper frequently invents entire passages of text when presented with moments of silence.

By [Wes Davis](#), a weekend editor who covers the latest in tech and entertainment. He has written news, reviews, and more as a tech journalist since 2020.

Oct 28, 2024, 12:19 AM GMT+1

[Link](#) [Facebook](#) [Twitter](#) | [27 Comments \(27 New\)](#)

A few months ago, my doctor showed off an AI transcription tool he used to record and summarize his patient meetings. In my case, the summary was fine, but researchers cited by *ABC News* have found that's not always the case with OpenAI's Whisper, which powers a tool many hospitals use — sometimes it just makes things up entirely.

Whisper is used by a company called Nabra for a medical transcription tool that it estimates has transcribed 7 million medical conversations, according to *ABC News*. More than 30,000 clinicians and 40 health

<https://www.theverge.com/2024/10/27/24281170/open-ai-whisper-hospitals-transcription-hallucinations-studies>









More AI devices making it through approval, but what about to the clinic?



DOI: [10.1056/Aloa2300030](https://doi.org/10.1056/Aloa2300030)

ORIGINAL ARTICLE

Characterizing the Clinical Adoption of Medical AI Devices through U.S. Insurance Claims

Kevin Wu , M.S.,¹ Eric Wu , M.S.,² Brandon Theodorou ,³ Weixin Liang , M.S.,⁴ Christina Mack , Ph.D.,⁵ Lucas Glass , Ph.D.,⁵ Jimeng Sun , Ph.D.,^{3,6} and James Zou , Ph.D.^{1,2,4}

Received: July 9, 2023; Revised: September 15, 2023; Accepted: September 29, 2023; Published: November 9, 2023

Abstract


There are now over 500 medical artificial intelligence (AI) devices that are approved by the U.S. Food and Drug Administration. However, little is known about where and how often these devices are actually used after regulatory approval. In this article, we systematically quantify the adoption and usage of medical AI devices in the United States by tracking Current Procedural Terminology (CPT) codes explicitly created for medical AI. CPT codes are widely used for documenting billing and payment for medical procedures, providing a measure of device utilization across different clinical settings. We examined a comprehensive nationwide claims database of 11 billion CPT claims between January 1, 2018, and June 1, 2023 to analyze the prevalence of medical AI devices based on submitted claims. Our results indicate that medical AI device adoption is still nascent, with most usage driven by a handful of leading devices. For example, only AI devices used for assessing coronary artery disease and for diagnosing diabetic retinopathy have accumulated more than 10,000 CPT claims. Furthermore, we found that zip codes that had a higher income level, were metropolitan, and had academic medical centers were much more likely to have medical AI usage. Our study sheds light on the current landscape of medical AI device adoption and usage in the United States, underscoring the need to further investigate barriers and incentives to promote equitable access and broader integration of AI technologies in health care.

How many AI algorithms/devices are reimbursed (in the USA) out of the 1000 devices approved by the FDA?

We need more randomized clinical trials for medical AI.

<https://onpub-media.nejmgroup-production.org/ai/media/b35da8b4-b078-492b-ae20-bf938063e91f.pdf>

The Business Imperative



By 2026, organizations that operationalize AI **transparency, trust** and **security** will see their AI models achieve a **50% result improvement** in terms of adoption, business goals and user acceptance.

Source: [Gartner](#)

What is Trustworthy AI?



Developing and using AI technologies in an **ethical** manner



Ensuring AI does not harm people



Asking not just, “Could we?”, but also “**Should we?**”



Building AI that reflects **our values** as a society

Principle-Driven Approach

Human-Centricity



Promote human **well-being**, human **agency** and **equity**.

Inclusivity



Ensure **accessibility** and include **diverse perspectives** and **experiences**.

Accountability



Proactively identify and mitigate adverse impacts.

Transparency



Explain and **instruct** on usage openly, including potential risks and how decisions are made.

Robustness



Operate **reliably** and **safely**, while enabling mechanisms that assess and manage potential risks throughout a system's lifecycle.

Privacy & Security



Respect the privacy of data subjects.

Oversight

AI Governance, Strategy and Enforcement

Operations

SOPs with Supporting Infrastructure

Compliance

Performance and Risk Management

Culture

Ethically Systemic Norms and Practices

Data Management

Data Quality

Variable Metadata

Data Preparation

Data Asset Catalog

Explanation

Natural Language Explanation

Explainable ML

Counterfactual Explanation

Surrogate Model Interpretation

Causal Inference

Detection

Bias Detection

Fairness Assessment

Privacy & Security

Privacy Preservation

Model Security

Autonomy Preservation

Consent and Control

Mitigation

Bias Mitigation

Bias Prevention

Synthetic Data Generation

Model Ops

Model Cards

Decisioning

Life Cycle Management

Metric Monitoring

Model Robustness

Model Oversight

GxP / pharma regulations

Patient Trust

HCP/HTA

Standards

Traceability

Technology

Trustworthy AI Landscape

ACTIVATION

Oversight

AI Governance, Strategy,
and Enforcement

Operations

SOPs w/ Supporting
Infrastructure

Data Management

Data
Quality

Variable
Metadata

Data
Preparation

Data Asset
Catalog

TECHNOLOGY

Detection

Bias
Detection

Fairness
Assessment

Mitigation

Bias
Mitigation

Bias
Prevention

Synthetic Data
Generation

DATA QUALITY

Includes features that address data quality concerns that have the potential to lead to pre-processing bias.

VARIABLE METADATA

Includes features that apply appropriate metadata and transparency principles to variables to ensure that downstream analytics can use this information for bias detection and mitigation.

DATA PREPARATION

Includes features that apply appropriate metadata and transparency principles to variables to ensure that downstream analytics can use this information for bias detection and mitigation.

DATA ASSET CATALOG

Includes features that allow a user to track data assets throughout the analytic lifecycle through the use of data sheets and lineage diagrams.

Trustworthy AI Landscape

ACTIVATION

Oversight
AI Governance, Strategy, and Enforcement

Operations
SOPs w/ Supporting Infrastructure

Accountability
Model Cards, Decisioning, Lifecycle Management, Metric Monitoring, Model Robustness, Model Oversight

TECHNOLOGY

Data Management

- Data Quality
- Variable Metadata
- Data Preparation
- Data Asset Catalog

Detection

- Bias Detection
- Fairness Assessment

Model Governance

- Natural Language Explanations
- Privacy Preservation
- Model Security
- Autonomy Preservation
- Consent & Control

Mitigation

- Bias Mitigation
- Bias Prevention
- Synthetic Data Generation

Model Ops

- Model Cards
- Decisioning
- Lifecycle Management
- Metric Monitoring
- Model Robustness
- Model Oversight

BIAS DETECTION
Includes features that will detect model performance/accuracy differences by various slices of the data. Also includes features to detect inappropriate model use.

FAIRNESS ASSESSMENT
Includes features that will allow users to assess the fairness of their models based on a range of fairness metrics/definitions.

Trustworthy AI Landscape

ACTIVATION

Oversight

AI Governance, Strategy,
and Enforcement

Operations

SOPs w/ Supporting
Infrastructure

Data Management

Data
Quality

Variable
Metadata

Data
Preparation

Data Asset
Catalog

TECHNOLOGY

Detection

Bias
Detection

Fairness
Assessment

Mitigation

Bias
Mitigation

Bias
Prevention

Synthetic Data
Generation

BIAS MITIGATION

Includes features that will allow the user to mitigate bias that was detected during model preparation, exploration, or analysis.

BIAS PREVENTION

Includes features that appropriately notify the user during the model building process to prevent unintentional introduction of bias.

SYNTHETIC DATA GENERATION

Includes features that allow the user to generate synthetic data using algorithms for bias remediation.

Trustworthy AI Landscape

ACTIVATION

NATURAL LANGUAGE EXPLANATION

Includes features that provide human-friendly explanations that can quickly give the user insights and information about their data and analyses.

EXPLAINABLE ML

Includes features that will allow a user to take advantage of machine learning models that are inherently explainable without the use of post-hoc surrogate model interpretation.

COUNTERFACTUAL ML

Includes features that allow the user to understand how model inputs impact the final output of a model.

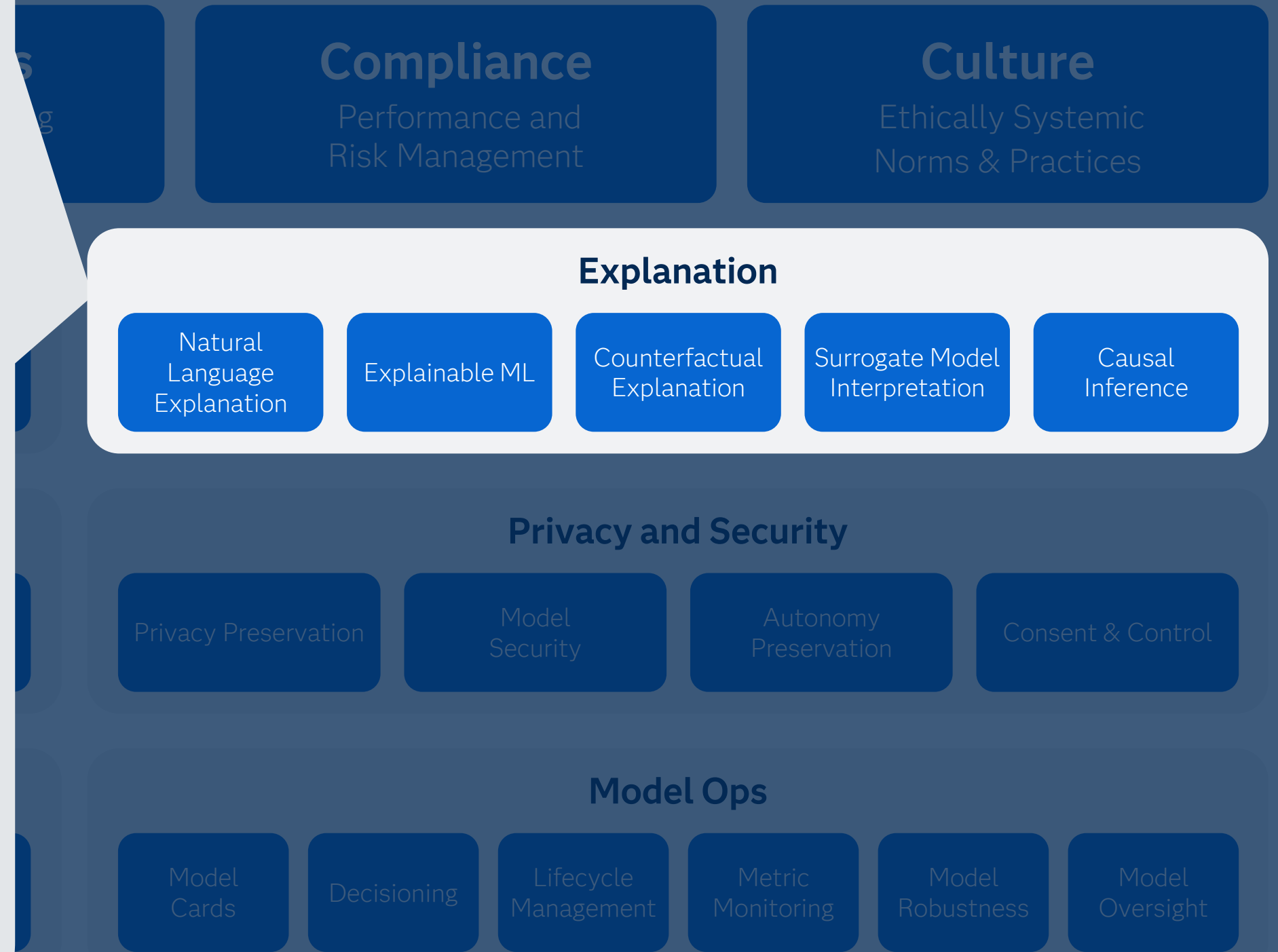
SURROGATE MODEL INTERPRETATION

Includes features that will allow a user to interpret black-box models that are not inherently explainable and to understand why/how an observation was assigned a given predicted probability.

CASUAL INFERENCE

Includes features that will allow a user to determine whether there is a causal (not just a correlative) relationship between two factors.

TECHNOLOGY



Trustworthy AI Landscape

ACTIVATION

TECHNOLOGY

PRIVACY PRESERVATION

Includes features that will help protect the privacy of those included in utilized datasets.

MODEL SECURITY

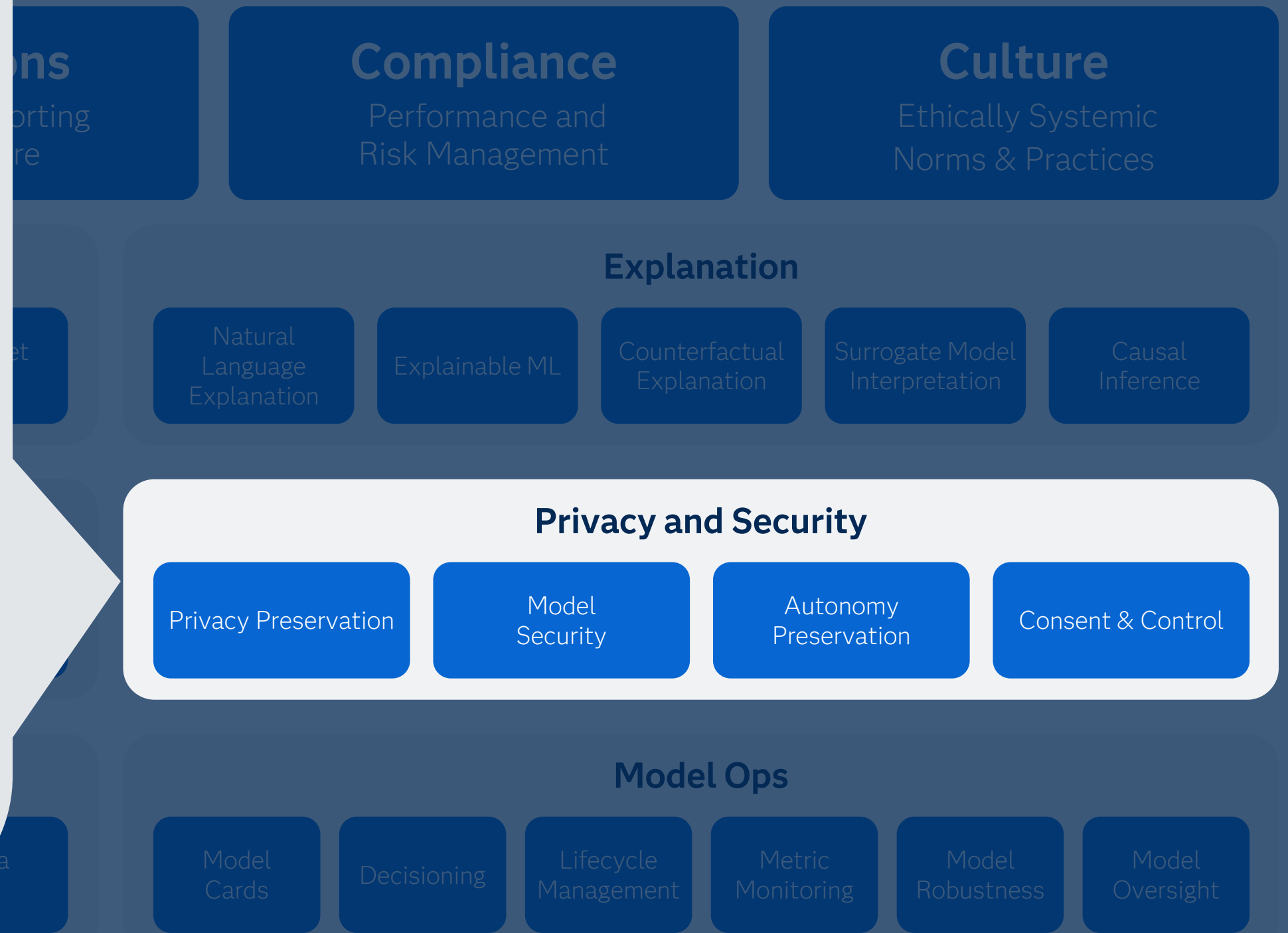
Includes features that will protect the model from intentional or unintentional misuse.

AUTONOMY PRESERVATION

Offer a mechanism for an authorized user to override or stop a data-driven system when needed.

CONSENT & CONTROL

Offer mechanisms for data subjects to be aware of how their data is used (including how data is used by third-parties) and offer a way for them to opt-out or consent.



MODEL CARDS

Includes features that allows users to document the intended use case, performance, limitations, model information/settings

DECISIONING

Includes features that allow users to document the business rules that need to be applied for decision making as well as how a particular model/set of models is used for a business use case; to justify the use of a model for decision making (versus a human-based decision).

LIFECYCLE MANAGEMENT

Includes features that will provide mechanism through which users will govern deployed models.

METRIC MONITORING

Includes features that will monitor metrics of interest in deployed models.

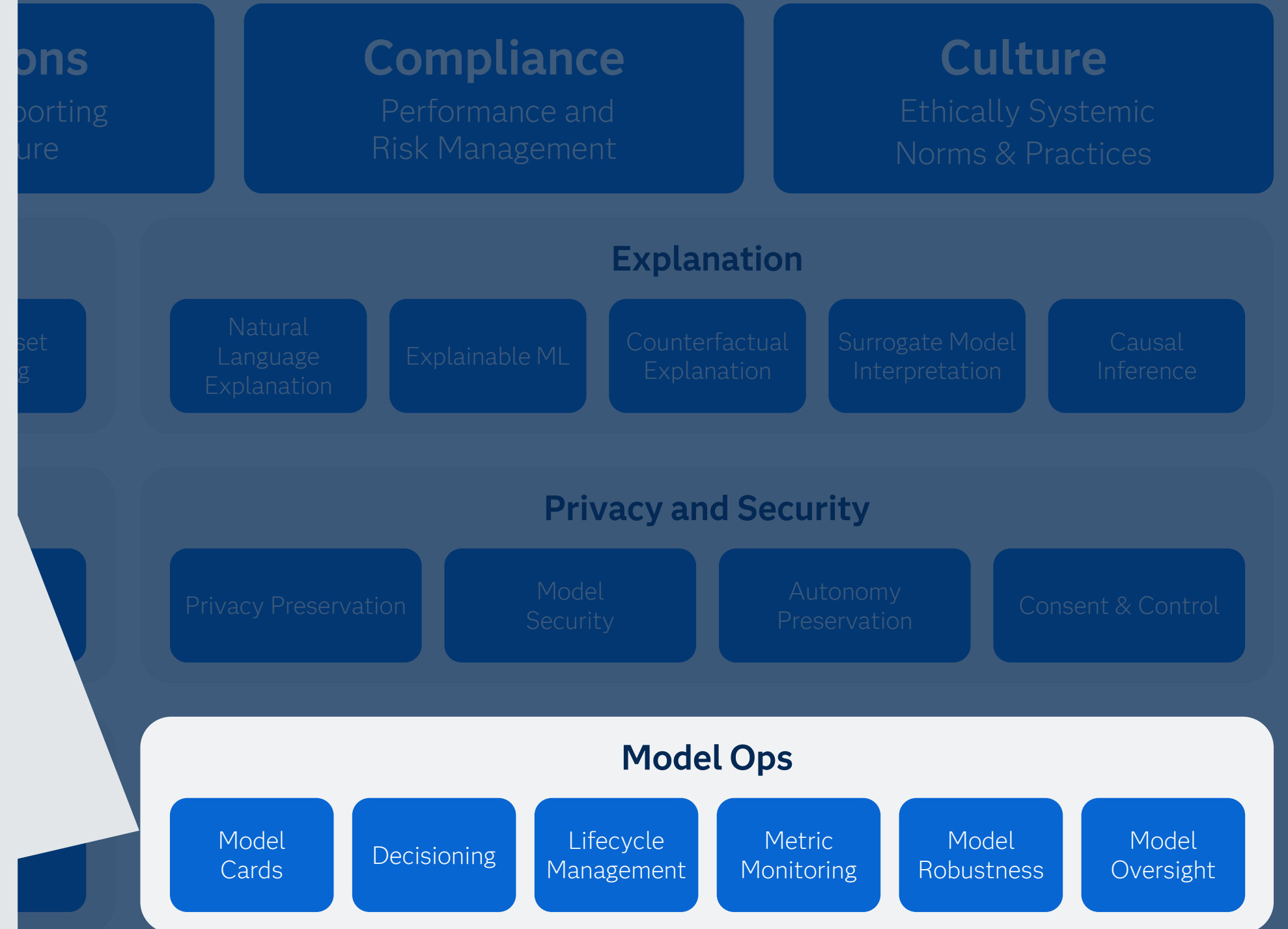
MODEL ROBUSTNESS

Includes features that will allow the user to test models to ensure they are robust and aren't susceptible to slight changes in input data.

MODEL OVERSIGHT

Includes features that will allow help users assess their models for compliance.

Model Ops Landscape



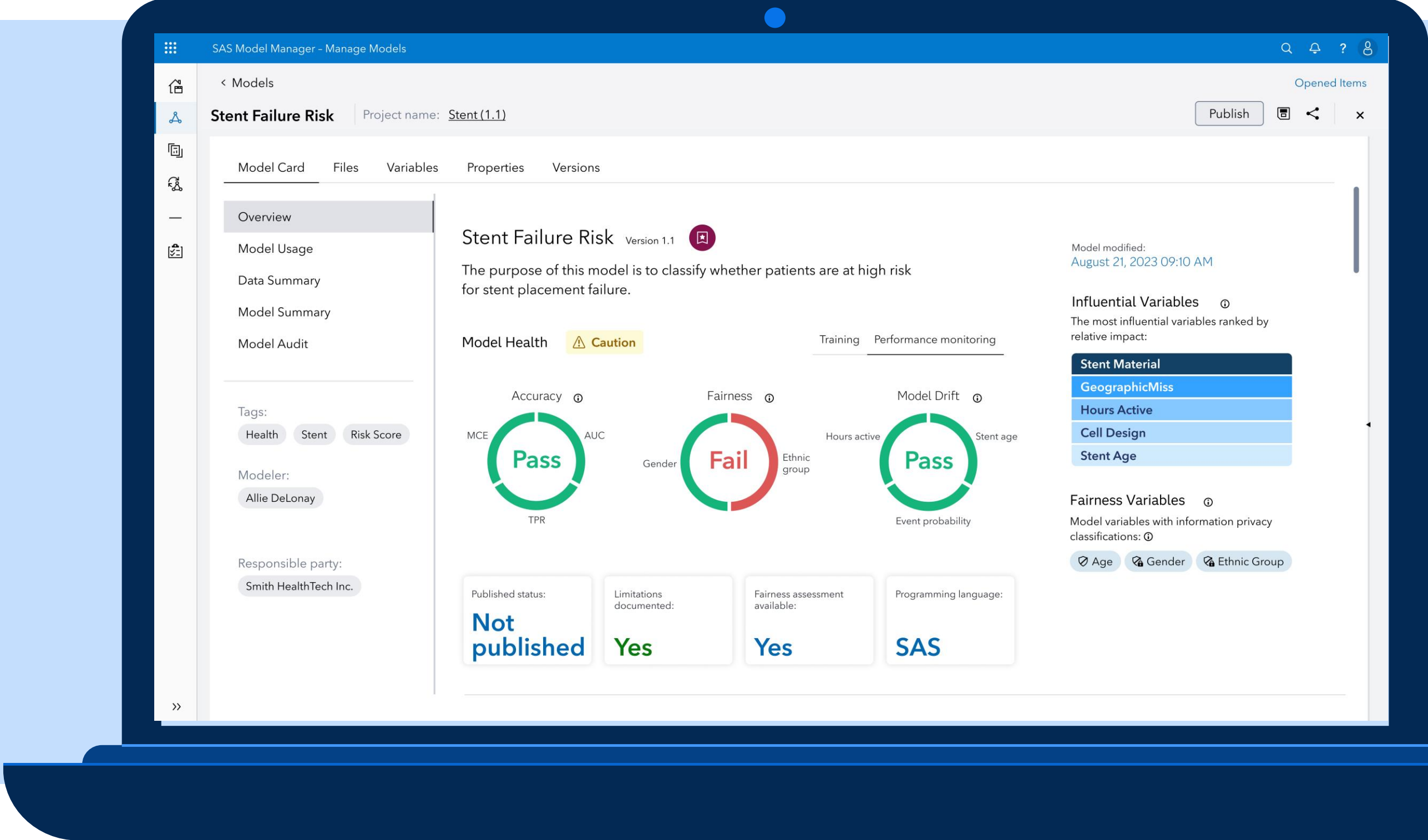
ACTIVATION

TECHNOLOGY

What are Model Cards?

“Nutrition labels for models” provide critical insight

Model cards generate a summary of a model’s training data, intended use and performance - demonstrating whether models are accurate, transparent and fair.



Fosters Trustworthy AI ● Provides Consumable Insights ● Enables Ethical Decision-Making





Stent Failure Risk

Project name: Stent (1.1)

Publish



Model Card

Files

Variables

Properties

Versions



Overview

Model Usage

Data Summary

Model Summary

Model Audit



Tags:

Health

Stent

Risk Score

Modeler:

Allie DeLonay

Responsible party:

Smith HealthTech Inc.

Stent Failure Risk Version 1.1

The purpose of this model is to classify whether patients are at high risk for stent placement failure.

Model Health **Caution**

Training | Performance monitoring

Accuracy ⓘ



MCE

Generalizability ⓘ



MCE Difference

Fairness ⓘ



Results are calculated with default thresholds. Modify your settings on the Properties tab of the project that contains this model for personalized results.

Project status:

Not published

Limitations documented:

Yes

Fairness assessment available:

Yes

Training code type:

SAS

Model modified:

August 21, 2023 09:10 AM

Influential Variables ⓘ

The most influential variables ranked by relative impact:

Stent Material
GeographicMiss
Hours Active
Cell Design
Stent Age

Fairness Variables ⓘ

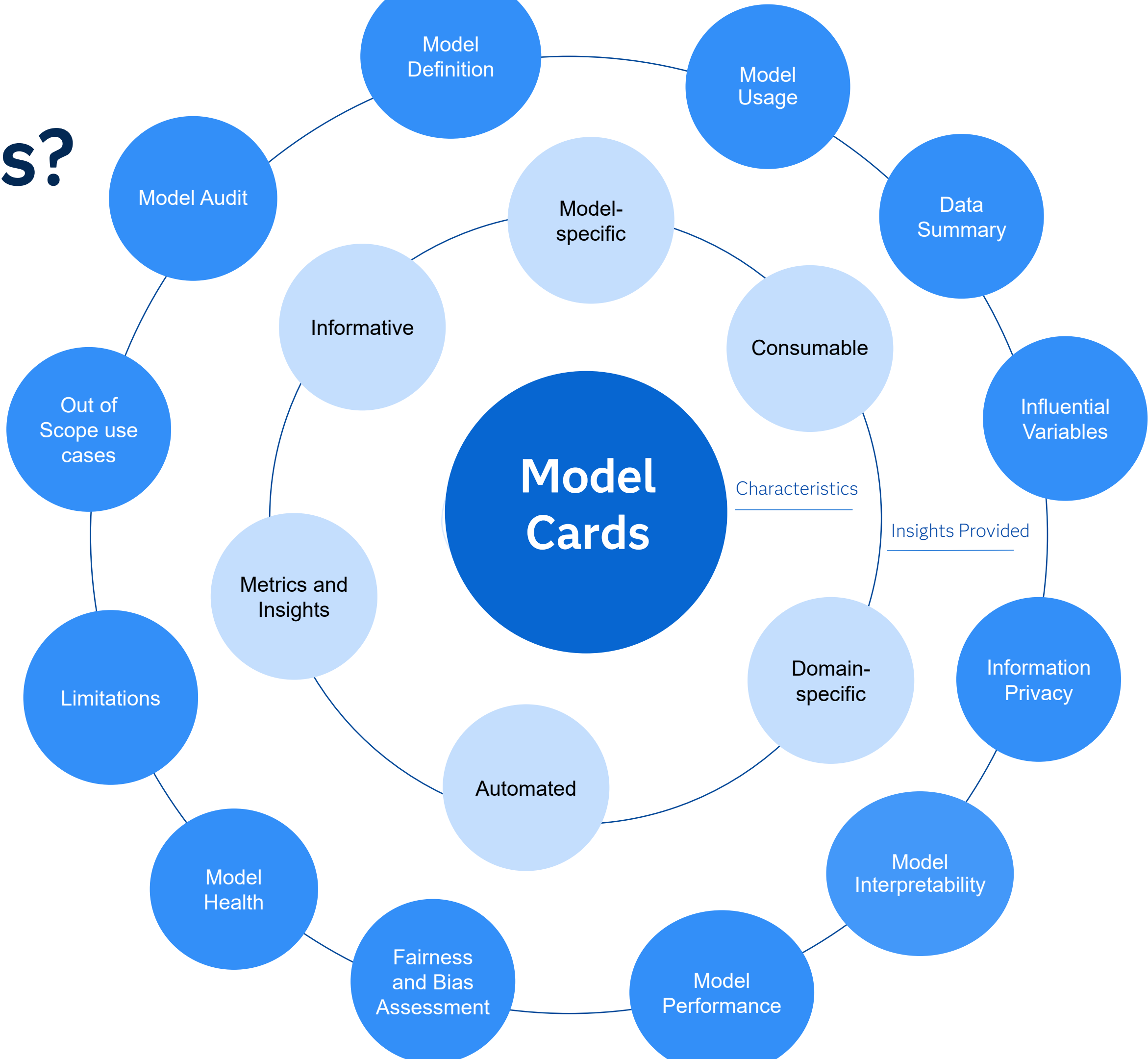
Model variables with information privacy classifications: ⓘ

- Age
- Gender
- Ethnic Group



What are a Model Cards?

Characteristics and Insights



Overview Tab

Key Insights and Use Cases

Summary of model performance, reliability compared to selected metrics and key influential variables.

Helps decide whether to put the model in production.

The screenshot displays the SAS Model Manager interface for a model named "Stent Failure Risk" (Version 1.1). The interface is divided into several sections:

- Navigation:** A left sidebar contains icons for home, search, and a list of models. The main area has tabs for "Model Card", "Files", "Variables", "Properties", and "Versions".
- Model Card:** The "Overview" tab is selected. It shows the model name, version, and a brief description: "The purpose of this model is to classify whether patients are at high risk for stent placement failure." It also indicates the model's health as "Caution".
- Model Health:** Three circular gauges are shown: "Accuracy" (Pass), "Fairness" (Fail), and "Model Drift" (Pass). Each gauge is associated with specific metrics: Accuracy (MCE, AUC, TPR), Fairness (Gender, Ethnic group), and Model Drift (Hours active, Stent age, Event probability).
- Model Metadata:** Includes "Model modified: August 21, 2023 09:10 AM", "Modeler: Allie DeLonay", and "Responsible party: Smith HealthTech Inc.".
- Model Status:** A "Published status" box shows "Not published". Other boxes indicate "Limitations documented: Yes", "Fairness assessment available: Yes", and "Programming language: SAS".
- Influential Variables:** A list of variables ranked by impact: Stent Material, GeographicMiss, Hours Active, Cell Design, and Stent Age.
- Fairness Variables:** A list of variables with information privacy classifications: Age, Gender, and Ethnic Group.

Model Usage Tab

Key Insights and Use Cases

Document intended use of the model, appropriate and out of scope use cases.

Prevent misuse and promote responsible deployment by making users aware of inappropriate applications and the model's limitations.

The screenshot displays the SAS Model Manager interface for a model named "Stent Failure Risk". The interface is divided into a left sidebar and a main content area. The sidebar contains a navigation menu with options: Overview, Model Usage (selected), Data Summary, Model Summary, and Model Audit. Below the menu, there are sections for Tags (Health, Stent, Risk Score), Modeler (Allie DeLonay), and Responsible party (Smith HealthTech Inc.). The main content area is titled "Model Usage" and contains three sections: "Intended Usage", "Expected Benefit", and "Out-of-Scope Use Cases".

Intended Usage
The risk score is calculated using information available at the time of surgery completion. If a patient is categorized as "high-risk" (score $\geq 26.7\%$), patients will have ischemic tests or catheterization scheduled; the appropriate follow-up time is to be determined by the physician. This score will help improve outcomes in high-risk patients while avoiding unnecessary, costly follow-up care for patients deemed to be "low risk".

Expected Benefit
The risk score will identify patients that will benefit from a telephonic intervention and therefore reduce the likelihood of stent failure.

Out-of-Scope Use Cases

- This model should not be applied to patients younger than 26.
- This model should not be used in new hospitals without additional testing and validation.

Limitations

- This model was fit using data from only four hospitals.
- Patient age was top-coded to 75, so it's poorly understood how accurate the model is for patients older than 75.
- There were fewer than 1,000 procedures that used a non-pharmacological stent coating or a biological stent coating which may result in less accurate results; if the patient's risk score is close to the threshold, consider classifying them as high-risk.

Data Summary

Key Insights and Use Cases

Analyze training data for insight into the AI's decision-making process.

Includes data source, data privacy considerations, data tags to indicate data types, data description, data relevance in training the model.

Includes analyses of data completeness, approval status and potential outliers.

The screenshot displays the SAS Model Manager interface for a model named "Stent Failure Risk". The interface includes a navigation sidebar on the left with options like Overview, Model Usage, Data Summary (selected), Model Summary, and Model Audit. The main content area shows the "Data Summary" tab with the following details:

- Columns: 20
- Rows: 22.9K
- Size: 6.5MB
- Status: Approved (with a green checkmark icon)
- Completeness: 80% (represented by a progress bar)

Additional information includes:

- Data last modified: August 21, 2023 09:10 AM
- Data last analyzed: August 21, 2023 09:10 AM
- Description: The dataset includes stent placement procedures at four hospitals in the UK; each observation is a unique patient procedure. Each stent placement procedure has a follow-up period of 12 months to track stent placement failure. Apart from the stent failure indicator, the remaining information was captured at the time of the procedure in the Electronic Health Record; this includes data about the patients, the procedure, and the materials used during the procedure. Given the data was gathered from four hospital EHRs, we identified formatting differences and standardized the variables prior to combining the data from each hospital.
- Tags: Health, Stent, Risk Score
- Modeler: Allie DeLonay
- Responsible party: Smith HealthTech Inc.
- Data Tags: EHR, Procedures, UK
- Training Data: STENT_FAILURE.sashdat
- Storage format: CAS
- Outliers: The data contains outliers and has values that could be considered private or sensitive.

An "Information Privacy" section is also visible, showing variable classifications:

- Private:** age_cat, missing_age_ind, age_cat_txt, Age, last_name, Name, nofare_ind
- Sensitive:** Gender
- Candidate:** No candidate variables detected in the data.

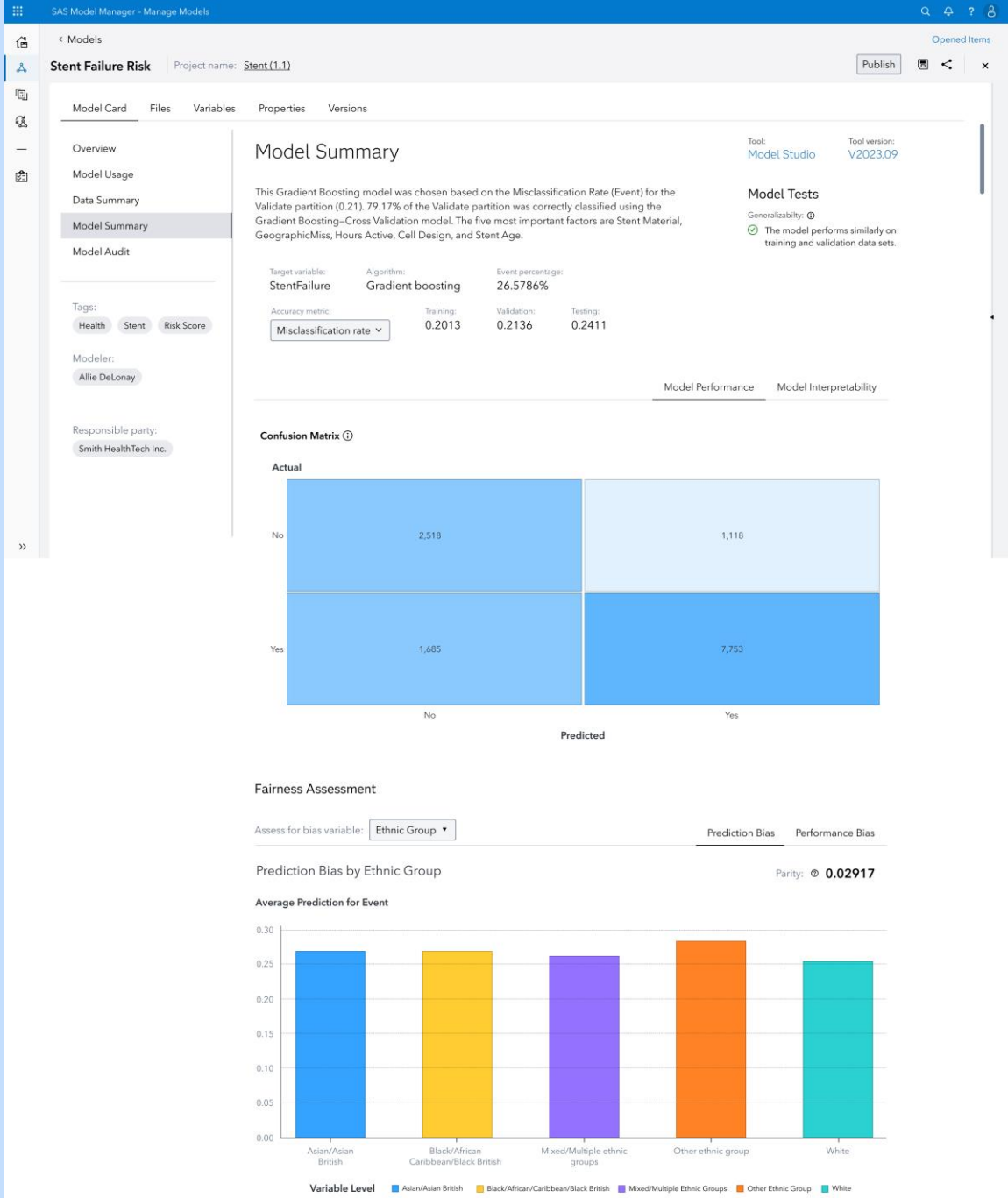
Model Summary

Key Insights and Use Cases

Understand granular model performance results, fairness assessment results, and generalizability metrics.

Review model interpretability, which indicates the variables most useful for model predictions.

Include governance details like the modeler, responsible party and model version.



Model Audit

Key Insights and Use Cases

Monitor an AI model over time to ensure accuracy, fairness, and avoid model drift.

Understand when the AI model was last evaluated, and which data was used for performance evaluation.

The screenshot displays the SAS Model Manager interface for a model named 'Stent Failure Risk'. The interface includes a navigation sidebar on the left with options like Overview, Model Usage, Data Summary, Model Summary, and Model Audit (which is currently selected). The main content area shows the Model Audit details, including performance monitoring information and a table of KPIs.

Model Audit Details:

- Project name: Stent (1.1)
- Performance Monitoring Last Run: August 22, 2023, 03:44 AM
- Project last run: August 30, 2023, 03:44 AM
- Data used: STENT_FAILURE_1_Q1
- Data used: STENT_FAILURE_3_Q3

KPI	Alert Condition	Alert Threshold	Value	Status
Accuracy				
Misclassification	Greater than	0.3	0.2	Pass
True positive rate	Less than	0.7	0.8	Pass
F1 score	Less than	0.8	0.9	Pass
Fairness				
Equal opportunity: Ethnic group	Greater than	0.1	0.2	Fail
Predictive parity: Gender	Greater than	0.1	0.05	Pass
Model Drift				
Input variable stability: Hours active	Greater than	0.2	0.15	Pass
Input variable stability: Stent age	Greater than	0.2	0.10	Pass
Output variable stability: Event probability	Greater than	0.3	0.25	Pass

Model Cards are an important element of your investment in Trustworthy AI



Trustworthy AI Workflow

Benefits

1

**Foster
Trustworthy AI
with 'at-a-glance'
view of the model**



2

**Consumable
insights accessible
to all personas
involved in
analytics process**



3

**Automatically
populate model
cards as the model
is developed,
managed and
deployed in SAS
Viya**



The AI toolbox of the pharmaceutical company

