

# Detect and Mitigate Ethical Risks



# Risk Management Process

Consists of identification, analysis, mitigation of risk in a cyclic process. So task of managing risk is never really done.

## Risk Identification

Start a **brainstorming session** with the project's key stakeholders. Each stakeholder should give their input about what risks they think the project will bring. For example, an IT might focus on the cyber-security vulnerabilities, whereas marketing focuses on the usability of the app for people with disabilities..

Go through the **key resources and documents**. Most projects have a list of requirements, meeting these requirements may introduce new types of risks. For example, a fitness app might be required to learn a user's daily exercise routine to help provide them with suggestions. Is there a privacy risk when it comes to the users health data?

Determine all the **factors** that are critical to the project's success. Then you look at each of these factors and think, how can it go wrong? What might its impact be?

Document your efforts, be sure to write down all the risks that were identified.

# Risk Management Process

**Risk Analysis** Risk analysis is about breaking down the risk into components to learn more about it.

One of the key goals of risk analysis is to discover a **risk's root cause**. For example, consider an app that learns about a user's exercise patterns. One potential root cause of a user's private health data being leaked is that the data was not stored in a confidential manner.

Risk analysis can also reveal **symptoms**, symptoms are the **manifestation of the root cause**. For example, a symptom might be that a malicious user is able to view the health data of another user without their consent.

## Qualitative Risk Analysis

Consider risk as a product of its likelihood to **occur** and its **impact**. Identify as **high, medium, and low**. So a risk might have a low likelihood of occurring, but a high impact if it does occur. It requires a lot of **subjective judgment**.

## Quantitative Risk Analysis

Use numbers to describe risk. Based on **statistical models** with lots of **historical data**. It will incorporate the project's cost, duration, the financial loss from risks that impacted the project and many more. **Not** all risk is so easily **quantifiable**, particularly risks involving third parties. **Complex** to calculate.

## Semi Quantitative Risk Analysis

Still uses terms like low and high to describe risk, but it also incorporates some statistical modeling to give you a more objective look at risk.

# Risk Management Process

**Risk Mitigation** Minimize the likelihood or impact of that risk. Eliminating risk entirely, is simply not possible.

Think in terms of **people, process, and technology**.

To mitigate the risk of malicious code execution in an app, you'll need to assign **personnel** to the task. These personnel might be professionals or beta testers.

Then you'll need to have some **process** in place to ensure that the people are able to do their work in a way that is repeatable and adds value to the business. For example, your process might be as follows for **App testers**; allocate the app and any hardware to the testers. Allow the testers to perform whatever activities they need to. Have the testers fill out a report after they're done, conduct a peer review of the reports. Processes enable people to do their jobs effectively.

Lastly you'll need **tools** to mitigate the risks. For example, the app testers might know what process to follow, but they need to be given the right testing tools.

Other options...

**Avoiding** risk, where you remove the activity or application that is causing the risk.

**Transferring** risk, where you offload the responsibility of addressing risk to a third party, i.e. Insurance company.

**Accepting** risk, where you determined that a risk is within your appetite and decide not to apply anymore mitigation tactics.

# Types of Ethical Risk





# Sources of Privacy Risks

## First-party data

Information collected **directly from customers or other target audiences**. It is obtained through surveys, feedback requests on websites or mobile apps, or in-store transactions etc. The data can be customer attitudes, contact information, social media interactions, and user behavior patterns. It provides insights into **user preferences, habits, and behaviors**. It serves various purposes including guiding marketing strategies, segmenting customer bases, personalizing user experiences, and more.

Users may **not fully understand** how their data will be utilized, leading to potential misconceptions. Additionally, users might **provide misleading information intentionally or inadvertently**.

## Third-party data

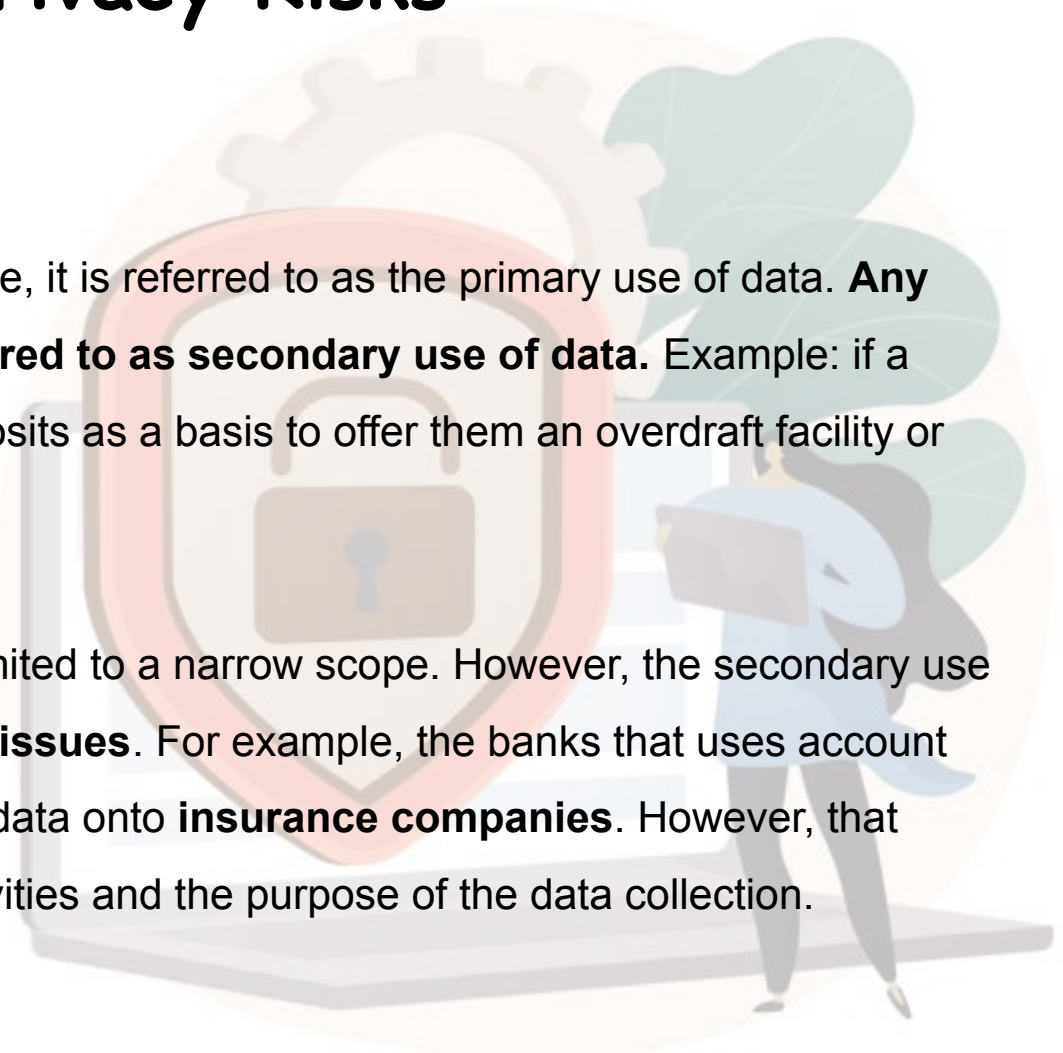
Collected from **one or more sources** without having a direct relationship with the users. Typically, multiple organizations are **aggregating** information. It can also refer to data about users that is **inferred rather than explicitly stated**, for example AI systems can make novel predictions about a person based upon existing data. Examples of third-party data include demographic data about customers such as age or gender. It carries **greater risks** compared to first-party data. The source of third-party data may be unknown, raising concerns about its ethical handling and consent. There's a risk of drawing false conclusions from inaccurate data, which may misrepresent individuals.

# Sources of Privacy Risks

## Secondary Use of Data

Whenever data is used in accordance with its purpose, it is referred to as the primary use of data. **Any use of data that falls outside of this intent is referred to as secondary use of data.** Example: if a bank uses data on a customer's regular account deposits as a basis to offer them an overdraft facility or personal loan, then that's probably legitimate.

When you use data for its primary purpose, you're limited to a narrow scope. However, the secondary use of data **opens up a much wider potential range of issues**. For example, the banks that use account deposit data to provide loans, could also **pass** the data onto **insurance companies**. However, that would be a major breach of the scope of agreed activities and the purpose of the data collection.



# Identifying Privacy Risks

01

Identify Personally Identifiable Information

02

Model Personas

03

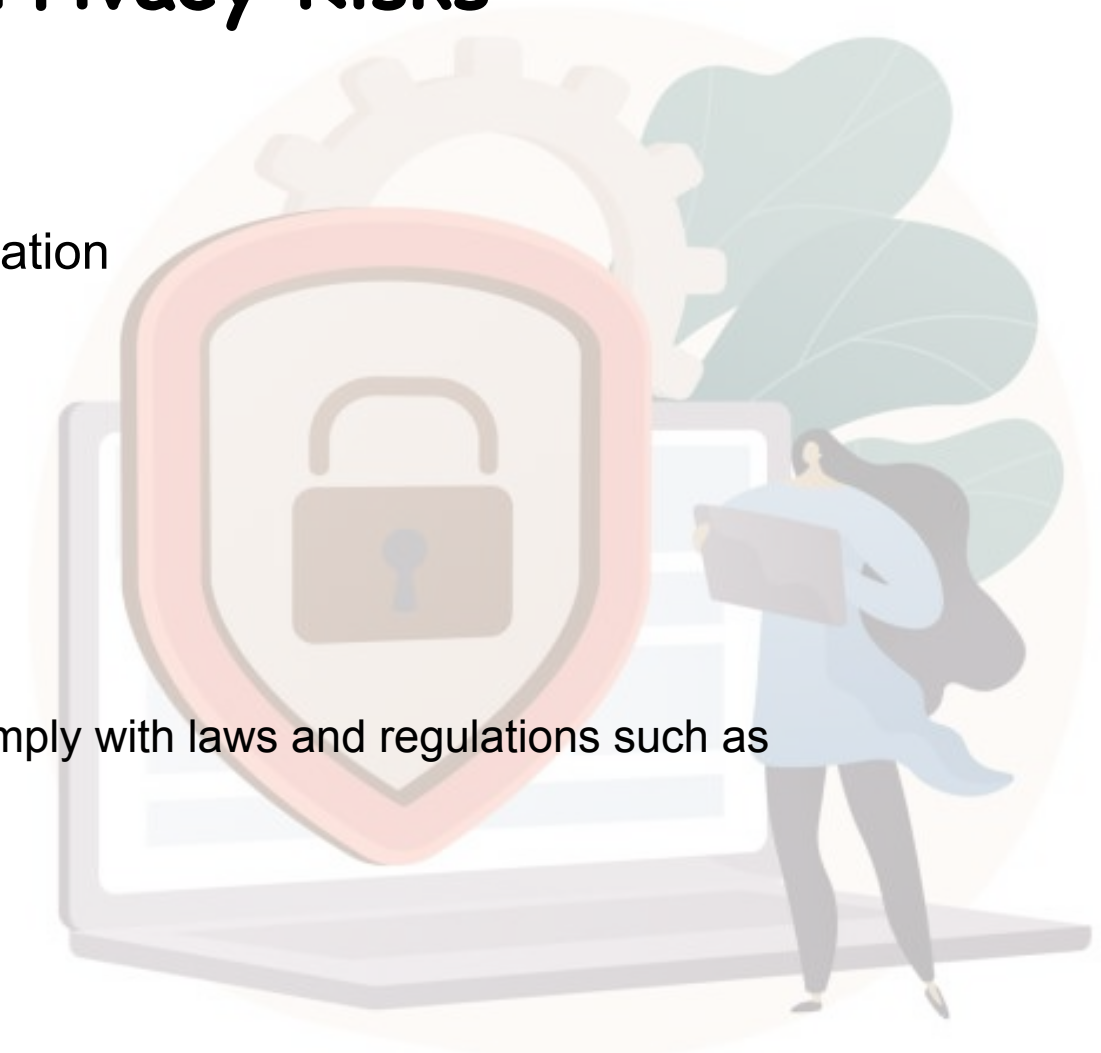
Track Customer Data

04

Meet Compliance Requirements - comply with laws and regulations such as GDPR or NIST Privacy Framework


05

Data Protection Policies





# Identifying Privacy Risks- Model Personas



JUSTINMIND


CHARLOTTE WALKER  
27, Los Angeles  
UX DESIGNER

· STATUS  
SINGLE


· SALARY  
\$50K

· TIER  
MID-LEVEL

· ARCHETYPE  
PERFECTIONIST


 PERSONALITY


- Prototyping
- Interviewing
- Design Thinking
- Empathy
- Coding


 BIO


Charlotte recently started a new job as a UX design in a mid-size bank. She moved over from the start-up world and is still getting used to all the changes, particularly the paperwork. She's excited to bring a user-focused perspective to the design department but nervous because she's the bank's first UXer.


Outside of the office she's a sports-mad psychology grad. She enjoys reading UX blogs and will sometimes go to UX-related conferences if they're nearby. She's also tuned into design channels like Dribbble.


 Motivations

IMPACT 


TEAMWORK 

PROMOTION 



USER NEEDS 

 Goals


- Introduce user focused mentality and methods into traditional company landscape
- Improve usability of bank's customer facing interfaces
- Grow the UX team


 Frustrations

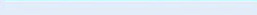
- Getting buy-in for the new department's activities
- Dealing with more bureaucracy than in her old job
- Communicating necessity for change to development team

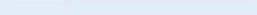
 "I want to help my team deliver great user experiences" 

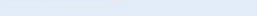
**Behavior**

Overseeing builds 

Writing specs 

Designing features 


Meetings 


User testing 


**Influences**

- CREDIBILITY
- COLLEAGUES
- TECHNOLOGY
- BLOGS/ FORUMS
- PSYCHOLOGY
- UI TRENDS

**Frequently used apps**

 Justinmind

 Google Calendar

 PocketGuard

# Mitigation Strategies for Privacy Risks

01

Intent & Consent

02

Minimize Private Data Sharing

03

Give the User Choices

04

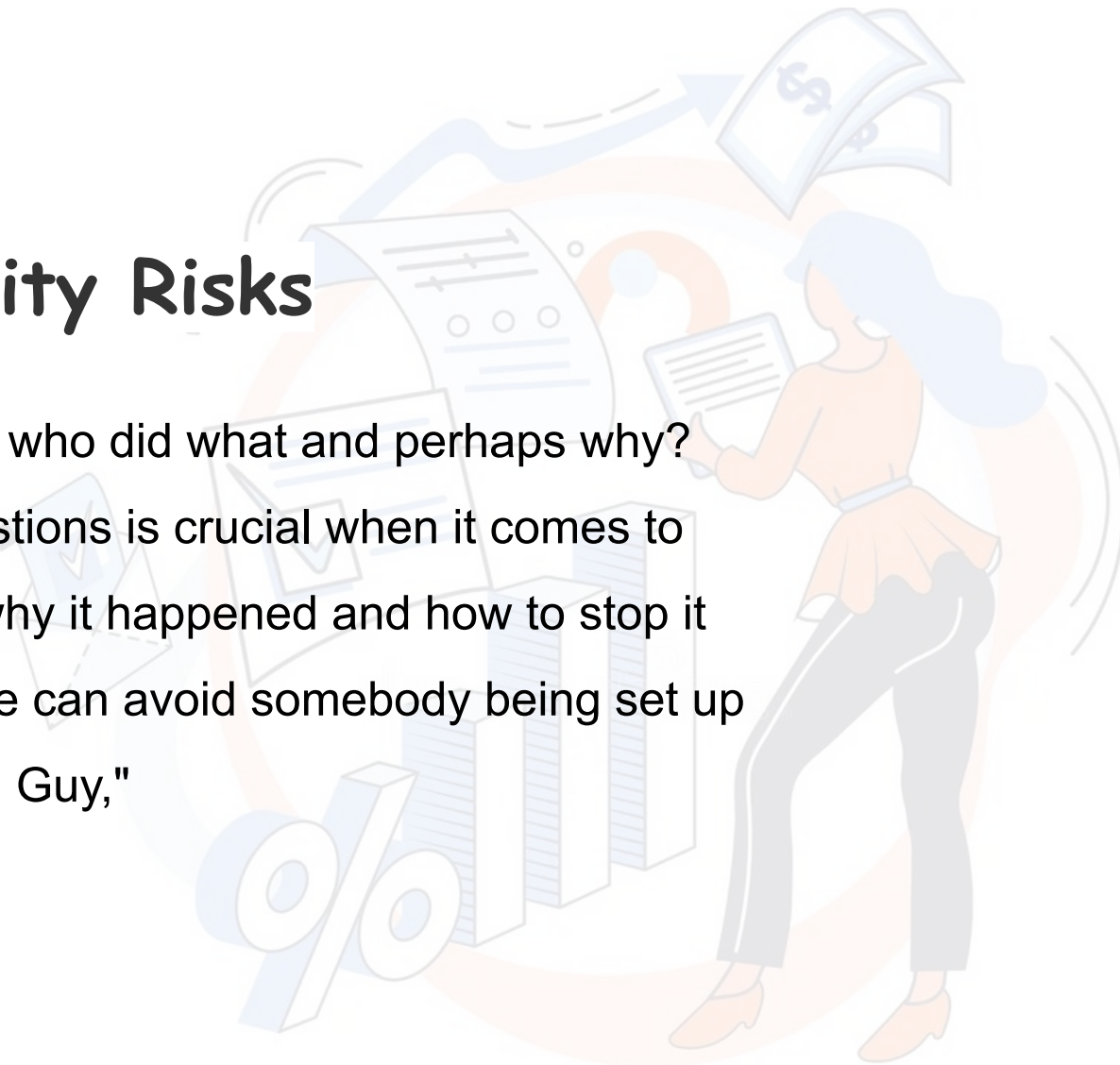
Minimize Private Data Collection



# Accountability Risks

Accountability, is all about understanding who did what and perhaps why?

Being able to answer those kinds of questions is crucial when it comes to investigating an incident, understanding why it happened and how to stop it from happening again. It also means that we can avoid somebody being set up to be "The Fall Guy,"



# Sources of Accountability Risks

## Use of Third-Party Components

Poses risks to accountability as organizations **rely on solutions they don't directly develop**, leading to a loss of control over various stages of the life cycle. Purchasing proprietary solutions can introduce challenges in assessing accountability due to closed-source nature, while open-source software, diffuses responsibility across multiple contributors, complicating accountability determination. Vulnerabilities discovered in widely used libraries highlights the complexities of assigning responsibility for ethical failures.

## Automation Bias

Leads to complacency and blind trust in machine-made decisions, even when contrary to common sense. It delegates control to automated systems, making it challenging to assign responsibility for decision-making failures, especially in complex AI-driven environments. Examples such as spell checkers, aircraft incidents, and autonomous vehicle accidents illustrate how automation bias can lead to serious consequences when human oversight is neglected.

## Extrajudicial Judgment

Poses a significant risk to **personal liberty** as decisions with serious consequences are made outside legal authority. Systems implementing such judgments, whether directly or indirectly, can lead to individuals being effectively banned from numerous establishments for minor infractions. The lack of recourse, transparency, and accountability amplifies the risks to individuals and society, highlighting the ethical responsibility of emerging technologists to consider the broader implications of their creations on liberty and justice.

## Lack of Guiding Principles

Without clear principles, it becomes ambiguous and challenging to define and address ethical failures, hindering accountability and fostering confusion about who should be held responsible. **Volkswagen** cars were programmed to give falsely low readings when they detected that they were being tested compared with normal use. Later, they had to pay more than 33 billion dollars in fines, recall costs and civil settlements as a result.

# Identifying Accountability Risks

## Recognize Black Box Algorithms

A black boxes' decisions are **difficult to interpret**, and therefore, upholding accountability becomes a challenge. We can either choose a path forward for the problem we're trying to solve, or we can choose to still use black box system, but we will be more prepared for any potential consequences. **Example:** Machine Learning algorithms.

## Assess the Organization's Governance Structure

It should also be structured so that the key players are held accountable whilst carrying out those tasks. It should also be structured so that these key players are held accountable whilst carrying out those tasks.



# Mitigation Strategies for Accountability Risks

01

Document and Distribute Company Policies

02

Document Design Processes

03

Document Auditing Processes

04

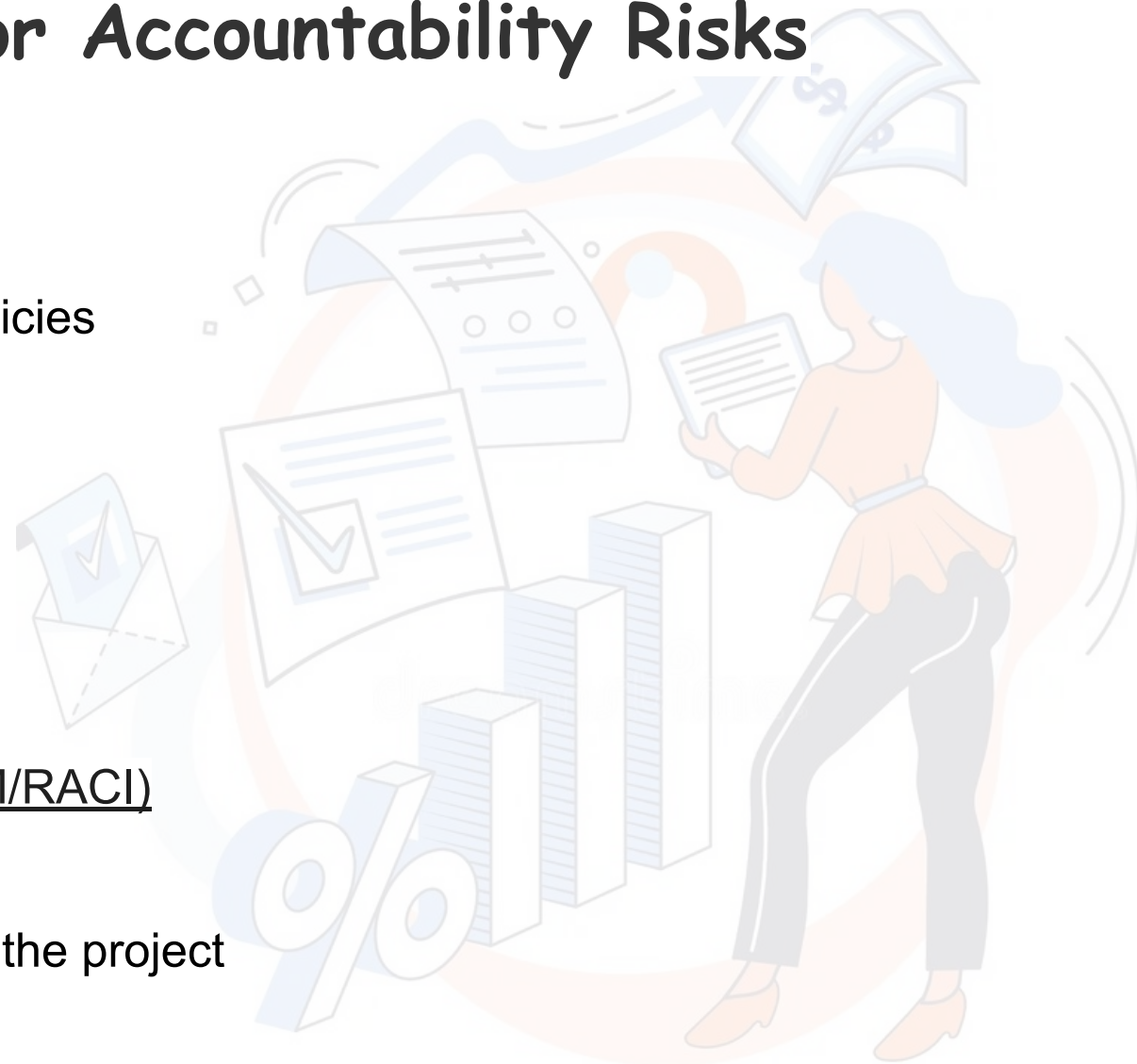
Responsibility Assignment Matrix (RAM/RACI)

05

Pilot Testing to assess the feasibility of the project

06

Collaboration with Data Sharing Partners





# Transparency and Explainability Risks

**Transparency:** Unrestricted access to information about processes, policies, and procedures related to technology.

**Explainability:** Clear explanation of why and how technology produces its results, addressing stakeholders' need to understand outcomes.



# Sources of Transparency and Explainability Risks

## Black Box Systems

Their inscrutable nature hinders the ability to verify model usefulness and understand decision-making logic, as seen in neural networks' hidden layers. This lack of transparency undermines user confidence, as it becomes challenging to assure them of the system's accuracy and freedom from bias,

## Third-Party Integration

When integrating third-party solutions into your data-driven projects, risks arise from insufficient visibility and transparency. For instance, if a third-party AI model lacks explainability, it becomes a black box within your system, hindering understanding and potentially creating a single point of failure. These risks can lead to difficulties in development, deployment, and maintenance, impacting the quality and timeliness of your projects and requiring careful management to mitigate.

## Intellectual Property Rights

Intellectual property rights, often closed source, can pose transparency risks in data-driven technologies, hindering independent auditing and compliance. For instance, closed-source AI code limits external verification of ethical practices, impacting user confidence, especially in sensitive areas like data privacy. Additionally, trade secrets integral to a product's function can further obscure transparency and explainability, complicating evaluation and usability for both customers and third parties.

## Shadow Banning

Shadow banning is banning or blocking someone on an online platform without that user's knowledge. Here, the banned user's comments and activity are not visible to the other users on the platform, but the banned user doesn't know this. However, it poses transparency & explainability risks. Users are unaware of their ban, undermining transparency, while automated systems may generate false positives, reducing user confidence and ethical concerns. Lack of explanation further exacerbates these risks, leading users to feel unfairly treated and diminishing trust in the platform.

# Sources of Transparency and Explainability Risks

## Black Box Systems



# Identify Transparency and Explainability Risks

## Explainable AI

Explainable AI is the opposite of a black box. The decisions that led up to the AI's outcome can be understood by a human. This leads to a greater degree of trust in those outcomes. Explainable AI can reveal risks because it can provide answers to the questions that black box systems raise. Some of these questions would include, why did an AI system make one decision and not another? Why did an AI system succeed or fail in a task? If there's a failure, how can we correct this failure?

## Identify Algorithmic Decisions

To proactively manage these risks, organizations must first identify when machine learning algorithms are making decisions versus human or traditional non-learning programs. Strategies such as clarifying AI goals during design and implementing oversight during deployment help distinguish algorithmic decisions, enhancing transparency and enabling effective management of transparency and explainability challenges.

## Deconstruct Specific Decisions

Deconstructing the decisions made by an AI model, such as a decision tree for predicting survival on the Titanic, reveals the logical progression of classifications and can aid in understanding model behavior. By tracing back through each decision node, we can discern the factors influencing predictions and gain transparency into the decision-making process. This transparency enables the identification of potential flaws or biases in the model's logic, supporting improvements and mitigating risks associated with transparency and explainability.

# Mitigation Strategies for Transparency and Explainability Risks

01

Explain How Systems Work

02

Help Users Seeking Explanations

03

Keep Humans in the Loop

04

Ensure Proper Data Disclosure

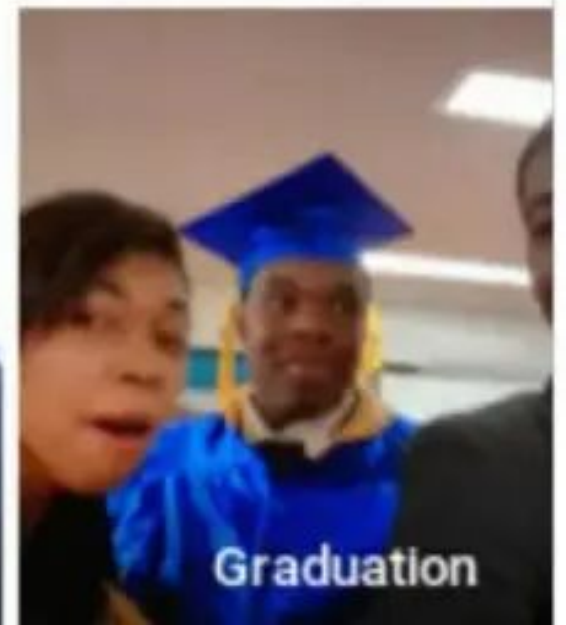
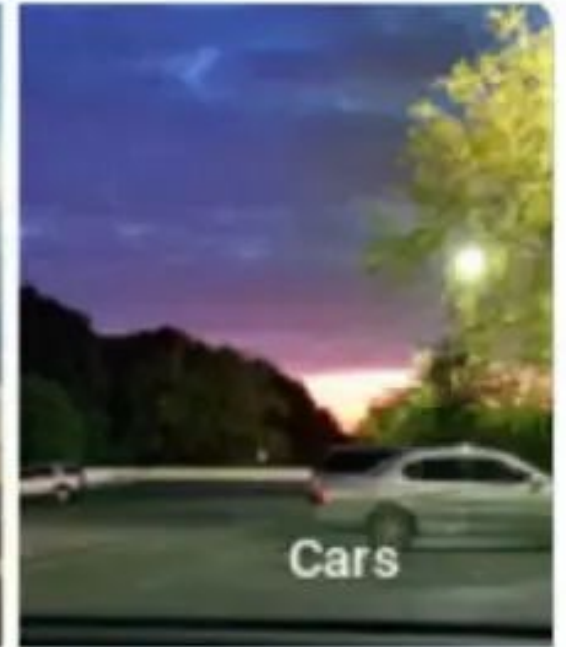
05

Be Upfront About Training Data Inadequacies



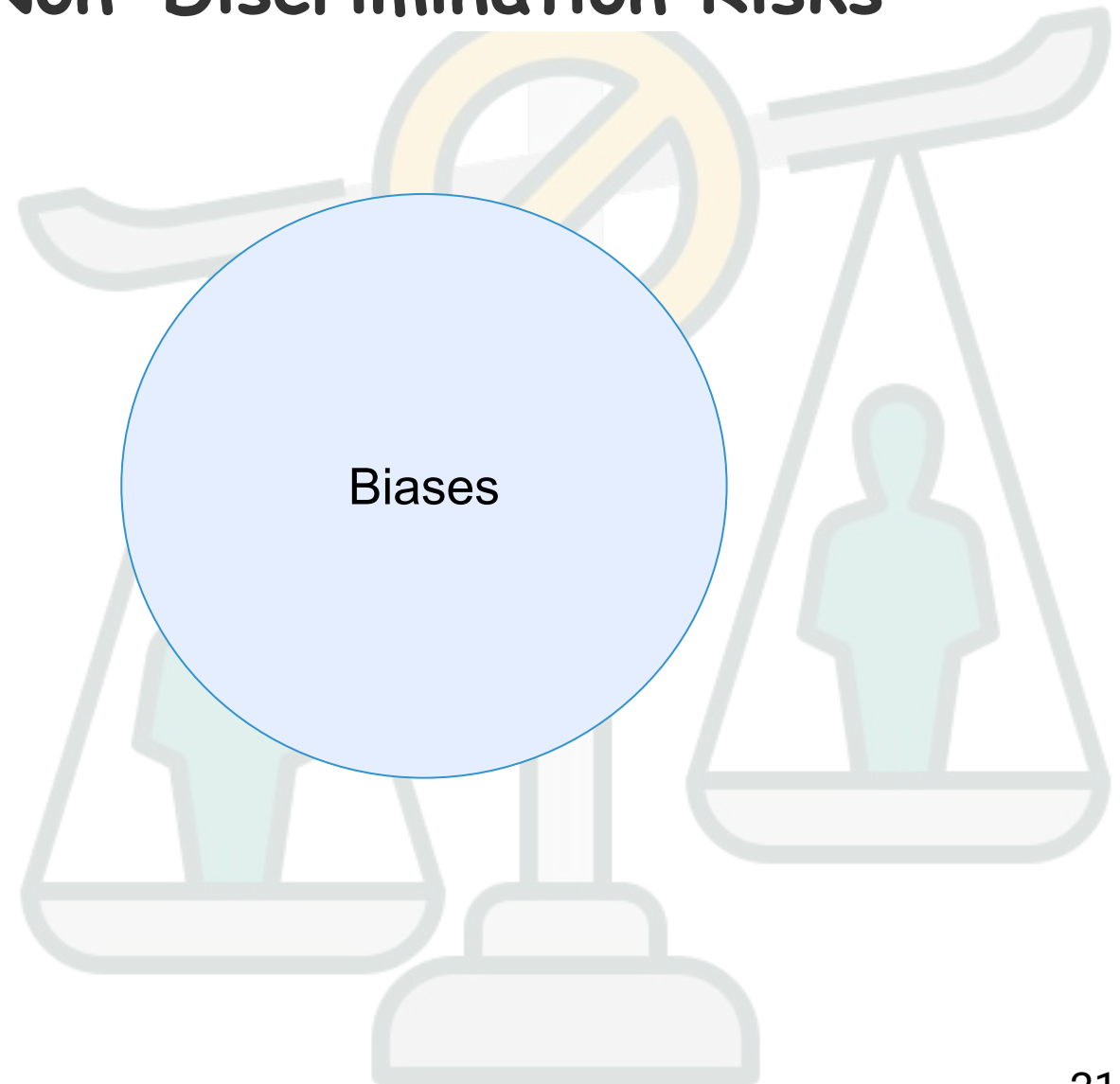
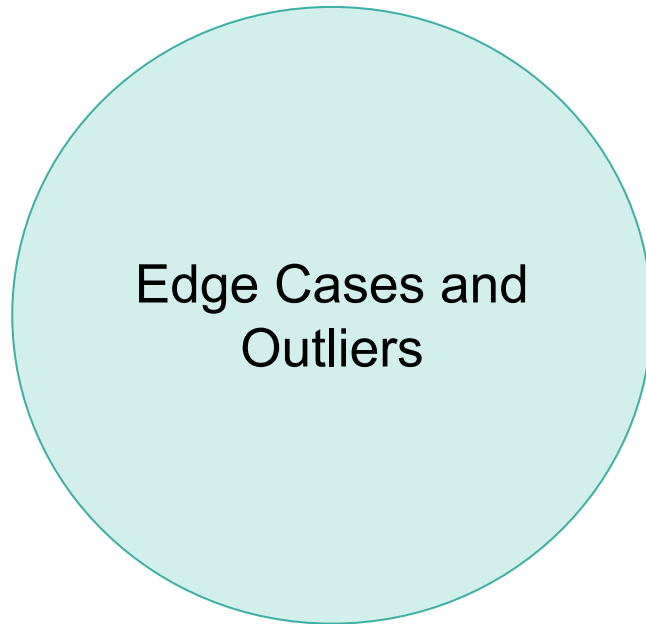
# Fairness and Non-Discrimination Risks

Google  
Photos





# Sources of Fairness and Non-Discrimination Risks



# Identify Fairness and Non-Discrimination Risks

## Analytical Techniques

A **systematic approach** to understanding data involves examining where it comes from and its key features. For instance, data collected directly from individuals may reveal sensitive information and impact their trust. In the case for **Surveys**, the methodology can impact what types of data, as well as how much data is collected.

It's important to consider how data was collected and whether it might be biased, such as if certain groups are overrepresented. Need to check whether the data is balanced or not.

**Exploratory Data Analysis (EDA)** helps you understand your data better by using techniques like summary stats and visual plots. It's important to do this before training a model to spot any issues or biases. EDA can show if there are gaps or problems in the data, so you can fix them before moving forward. For example, it might reveal outliers that could unfairly affect certain groups. Asking questions during data collection and training can also help uncover biases.

# Identify Fairness and Non-Discrimination Risks

## Analyze Models in Different Environments

When training and testing an AI model, it's crucial to consider the specific context and purpose. Even if the model seems unbiased in one scenario, changing the context can reveal hidden biases. For example, a recommender system might favor certain groups based on training data, but this might not represent the wider population accurately. Analyzing the model in different environments helps uncover biases and ensures fairness.

## Persona Modeling

Personas help identify potential biases in AI models by comparing them to the characteristics of real customers. The personas you create can point to large sections of your customer base who are members of protected groups. You can compare these personas to the model's results to see whether or not the model has any biases against these personas' representations or behavior. Likewise, you may model personas based on the available data and find that one race is represented much more commonly than another. This might indicate a deficiency in the data itself. Additionally, using AI to create personas can also reveal biases in the model itself. If the AI produces personas all from the same group, then the model itself might be impacted by biases during training, even if the training data isn't.

# Mitigation Strategies for Fairness and Non-Discrimination Risks

01

Pattern Matching vs. Bias

02

Inclusive Design and Foreseeability

03

Strategic analysis of external environments: STEEPV Analysis- social, technological, economic, environmental, political, and values

04

Perform User Testing

05

Gather Input from External Stakeholders