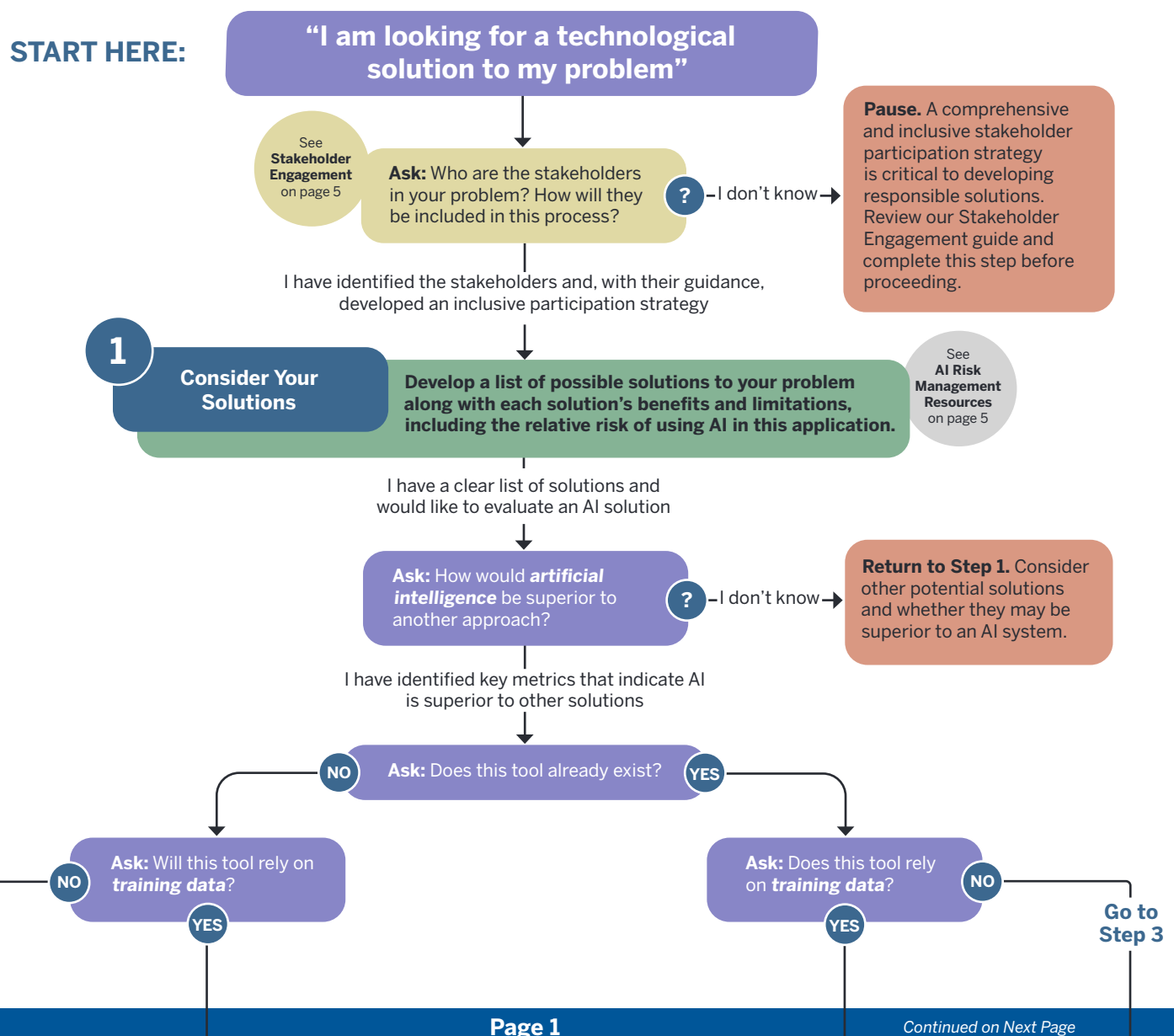


# Decision Tree for the Responsible Application of Artificial Intelligence (v1.0)

**BEFORE YOU BEGIN:** This decision tree is intended to be a guide that assists the user and their organization to structure the decision-making process on whether to develop or deploy AI solutions. Users should recognize that the model, while designed to be generally applicable, will not capture the nuances of every possible scenario. In any particular context, the answers may be difficult to judge, and despite the binary nature of the decision nodes, many questions may not have clear yes/no answers in real life. In such cases, collect as much information as possible, and use caution if deciding to proceed. The answers to many of these questions may be difficult to judge at the outset of a project, so this chart should be consulted periodically throughout the process of the development and deployment of an AI solution. Additionally, this tree is meant for AI solutions, but many of its core principles also apply to non-AI solutions.

Inclusive stakeholder engagement is central to this framework. Before applying the tree, consider who in your organization -or outside of it- would be best qualified to answer these questions. In most cases, the necessary skill set will be spread across multiple individuals. Always consider whether those providing responses are trustworthy, independent, and competent to do so. Assess the incentive structures that may be present for those providing input into decision-making, and at each step along the way, ask who would benefit and who might be harmed by the actions to be taken.

**\*\* Pages 5 & 6 contain the legend (color guide), definitions (of bolded italicized terms), and additional resources to accompany this decision tree. Familiarize yourself with those pages before you begin for optimal use.**



2

Consider the  
Training Data

**Ask:** What kind of **training data** will be required? Can they be sufficiently representative of the use case to solve the problem they intend to address? How do you know? Are you confident in this?

?

No/Unsure

Very confident

See  
Fundamental  
Rights  
on page 6

**Ask:** Can the **training data** be gathered in a way that respects ethics and human rights? How will you know? Should data depict victimization, can the rights and dignity of the victims be protected?

?

No/Unsure

Robust safeguards  
have been designedSee  
Harms from  
Automated  
Systems  
on page 6

**Ask:** Will the **training data** and their outputs be handled in a way that mitigates risks to participants? Consider data privacy.

?

No/Unsure

Data privacy and security  
safeguards are in place

**Ask:** Are the **training data** verifiable? Based on the answer, how confident can one be of their **accuracy**?

?

Not so  
confident

Very confident

**Ask:** Will the applicability of the **training data** to the problem change over time, or according to other variables like geography? Can these changes be addressed?

?

No/Unsure

There is a plan to  
account for this

**Develop procedures to  
periodically review relevance  
and applicability of training data.**

2

Consider the  
Training Data

**Ask:** What kind of **training data** were used to train the algorithm? Are the data applicable to my situation? How do you know? Are you confident in this?

?

No/Unsure

Very confident

See  
Fundamental  
Rights  
on page 6

**Ask:** Were the **training data** gathered in a way that respects ethics and Human Rights? Can you confirm this? Should data depict victimization, can the rights and dignity of the victims be protected?

?

No/Unsure

Robust safeguards  
were in placeSee  
Harms from  
Automated  
Systems  
on page 6

**Ask:** Are the **training data** and their outputs being handled in a way that mitigates risks to participants? Consider data privacy.

?

No/Unsure

Data privacy and security  
safeguards were in place

**Ask:** Are the **training data** verifiable? Based on the answer, how confident can one be of their **accuracy**?

?

Not so  
confident

Very confident

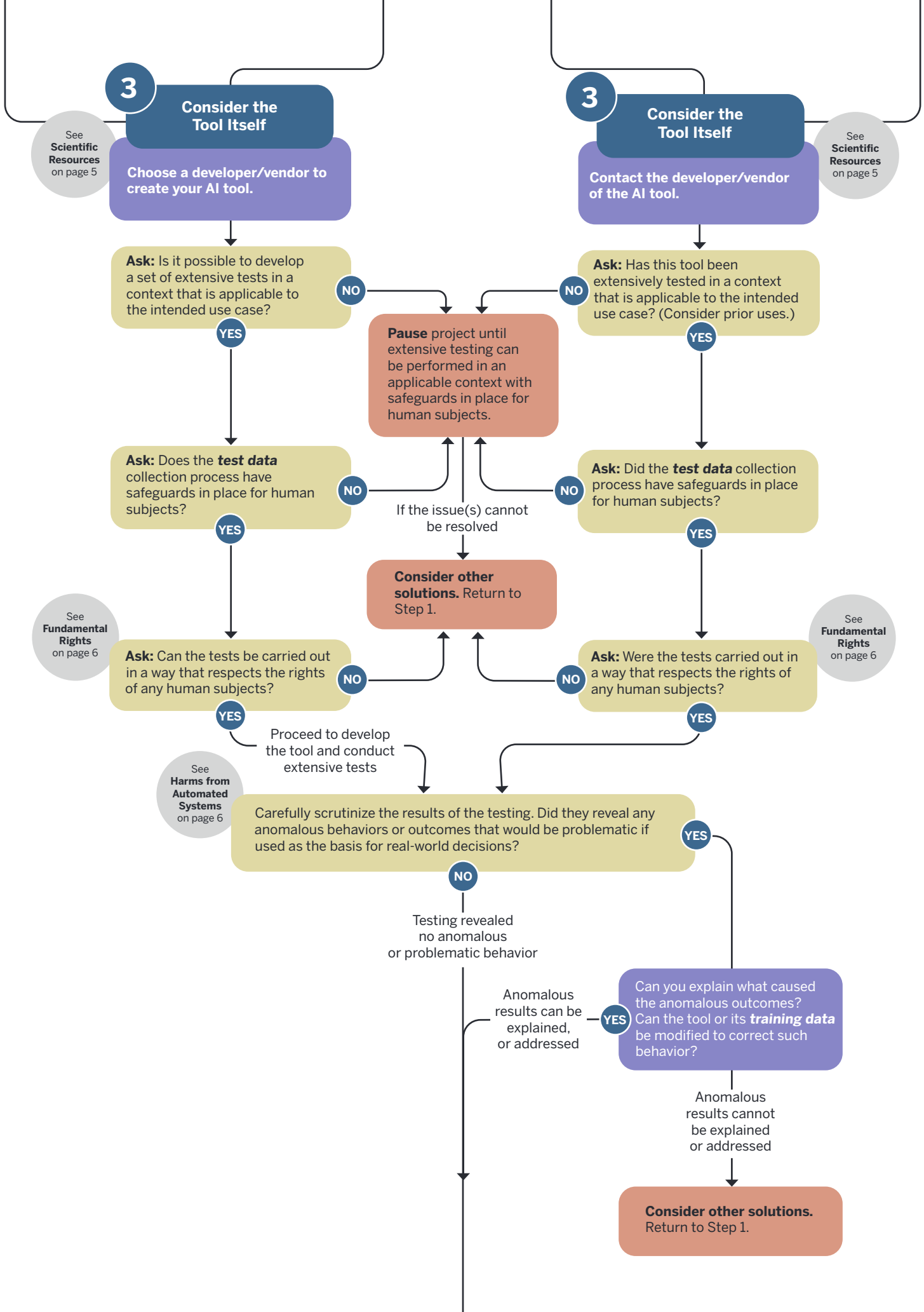
**Ask:** When and where were the **training data** collected? Are they still relevant to the problem to be solved? Will their relevance change over time, or according to other variables like geography?

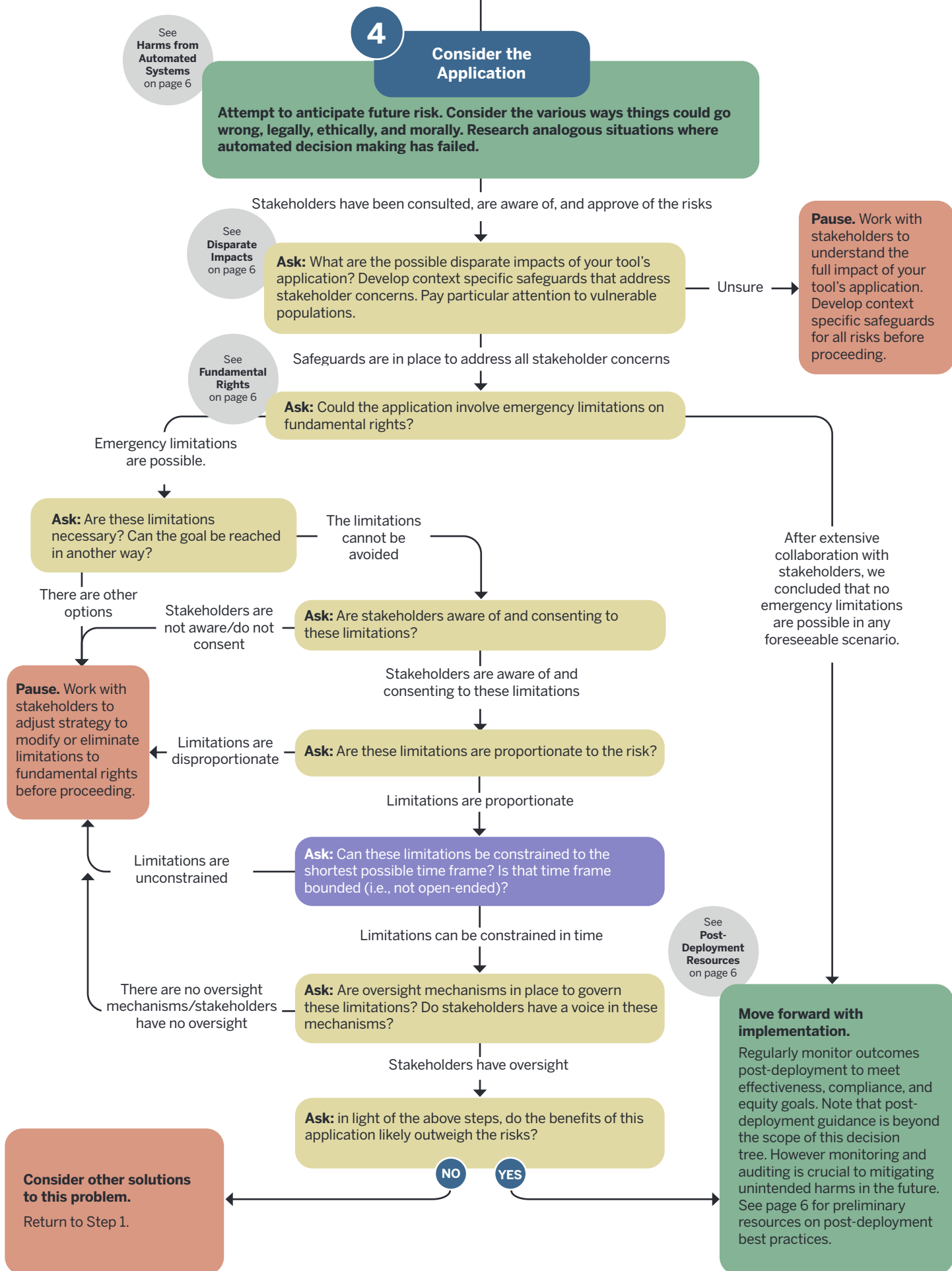
?

No/Unsure

There is a plan to  
account for this







**Develop procedures to  
periodically review relevance  
and applicability of training data.**





# RESOURCES accompanying the Decision Tree for the Responsible Application of Artificial Intelligence (v1.0)

## Decision Tree Legend

-  **You have reached a new step in the tree**
-  **Refer to the Resources for additional information**
-  **You should be consulting your stakeholders**
-  **Stop and follow the instructions before proceeding**
-  **Be mindful of these recommendations**
-  **You should follow the directions in the box**

## Stakeholder Engagement

Stakeholders are individuals and/or groups that are impacted by the problem you are hoping to address. These same stakeholders are likely to also be affected by the AI solution you are evaluating. The term “stakeholders” does not refer to a homogenous group. Stakeholders can include multiple individuals/groups with unique (and possibly conflicting) values and interests.

While we use the term “stakeholders” throughout the decision tree, different steps in the tree may involve different sets of stakeholders. When you are identifying the stakeholders, be sure to include all possible impacted individuals and/or groups and continuously review your list to reflect changes across time.

Partnership on AI (PAI)'s white paper “Making AI Inclusive” provides **four guiding principles** for ethical stakeholder engagement in AI/ML development:

1. *All participation is a form of labor that should be recognized*
2. *Stakeholder engagement must address inherent power asymmetries*
3. *Inclusion and participation can be integrated across all stages of the development lifecycle*
4. *Inclusion and participation must be integrated to the application of other responsible AI principles*

Additionally, PAI offers **three recommendations** aligned with these principles:

1. *Allocate time and resources to promote inclusive development*
2. *Adopt inclusive strategies before development begins*
3. *Train towards an integrated understanding of ethics*

Inspired by these principles and recommendations, this decision tree places inclusive stakeholder engagement at the center of the responsible AI framework. The yellow boxes refer to (suggested) points in the tree that call for stakeholder engagement. Once you have identified all the stakeholders, work together with them to create an **inclusive stakeholder engagement strategy** centered around stakeholder preference (e.g.: consider when they would - or would not - like to be consulted, via which channels of communication, with what kind of compensation, etc.).

For additional guidance, read the full PAI paper [here](#).

## AI Risk Management Resources

The National Institute of Standards and Technology (NIST) produced the AI Risk Management Framework (AI RMF 1.0) to be a resource for the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. The RMF describes seven traits of trustworthy AI as valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with their harmful biases managed.

The RMF outlines four functions for managing risk throughout the lifecycle of AI: mapping the context of the AI system and identifying associated risks; measuring those risks; managing risks through prioritization and regular monitoring; and governing to ensure compliance, evaluation, and leadership to cultivate a culture of risk management within an organization and decrease the likelihood of negative impacts.

For additional guidance, see the full NIST Framework [here](#).

The NIST AI RMF 1.0 acknowledges that risk cannot be eliminated, necessitating risk prioritization. AI systems that directly interact with or impact humans, or those which were trained on large datasets containing sensitive or protected information might warrant higher initial prioritization for risk assessment. Further consideration of risk prioritization can be found in the European Union's recent proposal on the regulation of AI, which introduces a tiered evaluation of risks:

- **Unacceptable risk:** AI systems considered a clear threat to the safety, livelihoods and rights of people.
- **High risk:** Included, but not limited to, technologies involving critical infrastructures, educational or vocational training, safety components of products, employment, management of workers and access to self-employment, essential services, law enforcement that may interfere with fundamental rights, migration, asylum and border control movement, and administration of justice and democratic processes.
- **Limited risk:** AI systems which should comply with minimal transparency requirements that would allow users to make informed decisions.
- **Minimal risk:** Systems with high transparency and minimal threats to people, such as AI-enabled spam filters.

For additional information, read the full policy [here](#).

## Definitions

The AAAS “Artificial Intelligence and the Courts: Materials for Judges” includes a Foundational Issues and Glossary that provides definitions for key terms (bolded-italicized) used in this decision tree:

**Accuracy:** The ability to produce a correct or true value relative to a defined parameter.

**Artificial Intelligence (AI):** No widely agreed upon definition. AI is both a concept and a category of technology tools that are powered by advanced mathematical models and data that can augment, replicate or improve upon the type of human cognitive task that otherwise requires thinking, beyond calculating.

**Test Data:** The data used to evaluate how well a trained model is performing once it is built and before it is released.

**Training Data:** The historical data used to develop and teach an AI model the logic and pattern recognition to generate desired predictions in the future.

For additional guidance, read the full paper [here](#).

## Scientific Resources

For general questions, contact AAAS at [srj@aaas.org](mailto:srj@aaas.org).

If you need explicit scientific guidance, reach out to On-Call Scientists at [oncall@aaas.org](mailto:oncall@aaas.org) and we will attempt to match you with an expert who can assist in your specific case. Additionally, consider reaching out to your local scientific community or university resources.

## Fundamental Rights

The AAAS Framework for the Responsible Development of AI Research identified four overall guiding principles for evaluating AI in the context of human rights.

These are:

- **Informed Consent** - individuals have autonomy to make their own choices about participation
- **Beneficence** - AI applications must benefit both the individual and the group
- **Nonmaleficence** - AI applications must not harm participants
- **Justice** - participants must be treated equally, and not be subject to disparate impacts

Additionally, the following rights must always be respected:

- Privacy
- Data Confidentiality
- Non-Discrimination
- Security of the person
- Freedom of Movement
- Freedom from being subjected to experimentation
- The right to enjoy the benefits of science
- Freedom from cruel, inhuman, or degrading treatment

For additional information, see the White House Blueprint for an AI Bill of Rights [here](#).

## Disparate Impacts

When evaluating AI applications in a human rights context, one should pay special attention to actions that may disproportionately impact marginalized and/or vulnerable populations.

In particular, consider the scale of the impact and the timing of the implementation. For example, are the tools deployed extensively? Are they already in use? Will they be soon? These factors may influence the presence or degree of disparate impact.

The following list of traits, although not exhaustive, provides examples of groups that are commonly subject to disproportionate impacts:

- Race and ethnicity
- Disability status
- Geographic location
- Immigration status
- Contact with the criminal justice system
- Socioeconomic status
- Dependence on safety nets
- Age
- Sexual orientation/Gender identity
- Religion, particularly if a minority

## Disclosure

\*This decision tree was funded by AAAS through the (AI)2: Artificial Intelligence — Applications/Implications initiative, which is supported by Microsoft. The interpretations and conclusions contained in this document are those of the authors and do not necessarily represent the views of the AAAS Board of Directors, its council and membership, or Microsoft.

## Types of Harm from Automated Systems

The same properties of speed, versatility, and flexibility that make AI a useful tool for automated decision making also have the potential to magnify the harm that such systems may cause when carelessly or improperly developed and deployed. By examining the harms that AI has been known to produce in previous circumstances, you can be better informed about the risks that may be present as you consider deploying AI tools in your own work. Throughout this process, keep in mind the precautionary principle: actions which present an uncertain potential for harm must be accompanied by measures to minimize the threat of that harm unless it can be shown that they present no appreciable risk.

This list outlines some common ways that AI can lead to harm. A more detailed elaboration of this list is available from Microsoft\* [here](#).

- Over-reliance on safety features
- Inadequate fail-safes
- Over-reliance on automation\*\*
- Distortion of reality or gaslighting
- Reduced self-esteem/reputation damage
- Addiction/attention hijacking
- Identity theft
- Misattribution
- Economic Exploitation
- Devaluation of Expertise
- Dehumanization
- Public Shaming
- Loss of Liberty
- Loss of Privacy
- Environmental Impact
- Erosion of Social & Democratic Structures
- Discrimination in: Employment, Housing, Insurance/Benefits, Education, Access to Technology, Credit, Access to / Pricing of Goods and Services

\*\*For more on Overreliance on AI, read the Microsoft report [here](#).

## Post-Deployment Resources

The resources below, though not exhaustive, are intended to serve as guidance for regular post-deployment monitoring and auditing of AI systems.

- “Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing,” [Raji et al.](#)
- “AI Auditing Framework,” [Institute of Internal Auditors](#)
- “AI Audit-Washing and Accountability,” [German Marshall Fund](#)
- “Using Algorithm Audits to Understand AI,” [Stanford University Human-Centered AI](#)

## Acknowledgements

The AAAS Center for Scientific Responsibility and Justice (CSRJ) would like to thank the following individuals for contributing their knowledge to this project. Without their invaluable insights, the creation of this tool would not have been possible.

Ronald Arkin  
Regents' Professor Emeritus,  
School of Interactive Computing,  
Georgia Institute of Technology

B Cavello  
Director of Emerging Technologies,  
Aspen Institute

Clarice Chan  
Former White House Presidential  
Innovation Fellow

Fredy Cumes

Michelle Ding  
Brown University  
AAAS CSRJ

Jonathan Drake  
AAAS CSRJ

Scott Edwards  
Program Director, Research and Advocacy,  
Amnesty International

Joel Ericson  
AAAS CSRJ

Juan E. Gilbert  
Andrew Banks Family Preeminence  
Endowed Professor & Chair,  
Computer & Information Science &  
Engineering Department,  
Herbert Wertheim College of Engineering,  
University of Florida

Danielle Grey-Stewart  
University of Oxford  
AAAS Center for Scientific Evidence in  
Public Issues

Theresa Harris  
AAAS CSRJ

Nick Hesterberg  
Executive Director,  
Environmental Defender Law Center

Jennifer Kuzma

Daniela Orozco Ramelli  
Senior Forensic Professional, EQUITAS

Jake Porway

Kate Stoll  
AAAS Center for Scientific Evidence in  
Public Issues

Mihaela Vorvoreanu  
Director, UX Research & RAI education,  
Aether, Microsoft