



Research, Applied Analytics, Statistics (RAAS)

Integrating Trustworthy Analytics

Jameson Carter, Chloe Zheng, Brian Xu

August 19, 2022



Our Team

Research, Applied Analytics, & Statistics (RAAS)

- Stephanie Needham, *Project Lead*
- Holly Donnelly, *Interim Project Lead*
- Alissa Graff, *Tools Workstream Lead*
- Frank Cousin, *Training Workstream Lead*
- Isaac Schwab
- Yan Sun, *Literature Review Workstream*
- Nathan Gire, *Literature Review Workstream*
- Traci Suiter
- Annelise Britten
- Tom Hertz
- Civic Digital Fellows (Summer 2022)
 - Jameson Carter
 - Chloe Zheng
 - Brian Xu

Equity, Diversity, & Inclusion (EDI)

- Jon Ocana
- Michael Sebastiani



What is Trustworthy Analytics?

Trustworthy Analytics is a manner of designing, developing, acquiring, and using AI tools and analytic tools in Public Algorithmic Systems in a way which **fosters public trust and confidence** while protecting privacy, civil rights, and civil liberties.

[HHS Trustworthy AI Playbook, p. 5, Pittsburgh Task Force on Public Algorithms, p. 8]



Project

Objective:

The objective of this research is to mitigate risk by recommending best practices form incorporating trustworthy analytics into RAAS analytic processes.

Research will include:

- Gaining an understanding of current private and public sector best practices.
- Identifying necessary data, testing tools, and methods.
- Increasing awareness throughout RAAS.

FY22 Activities:

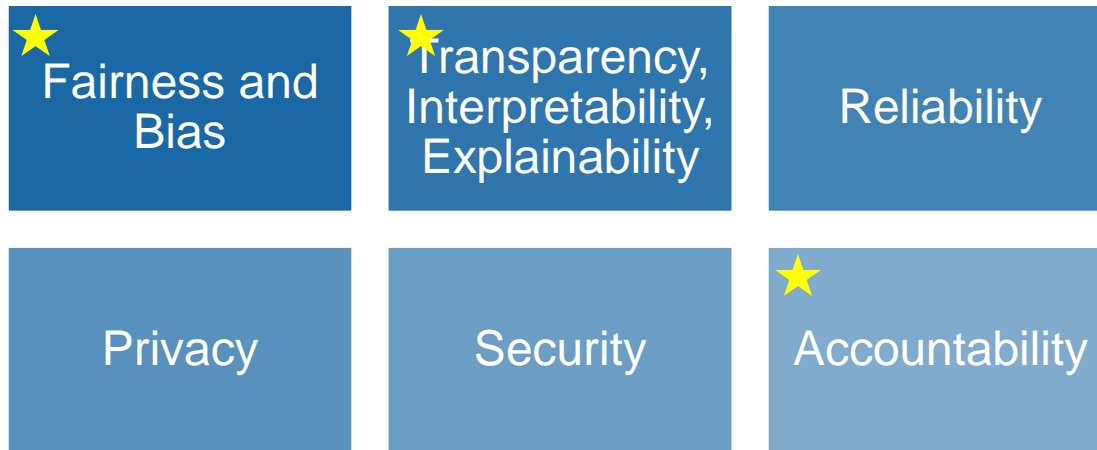
- **Conduct a literature review** to identify trustworthy analytics best practices and develop a trustworthy analytics toolkit to share best practices.
- **Conduct a pilot** to assess data, tools, and methods, and develop recommendations that should be considered and applied when developing and implementing analytic solutions.
- **Develop training solutions** on trustworthy analytics and deliver them to the RAAS Community.



Chloe: Overview of Principles

What does this cover?

We provide an overview of six principles that are imperative to foster and strengthen trust in analytic systems.



Why is this important?

As automation and artificial intelligence systems gain popularity, analytic teams must understand these principles – these ideas and concepts are recognized across international agencies, U.S. agencies, the private sector, and academic institutions. The overview covers risks and case studies that demonstrate consequences when teams fail to satisfy these principles.

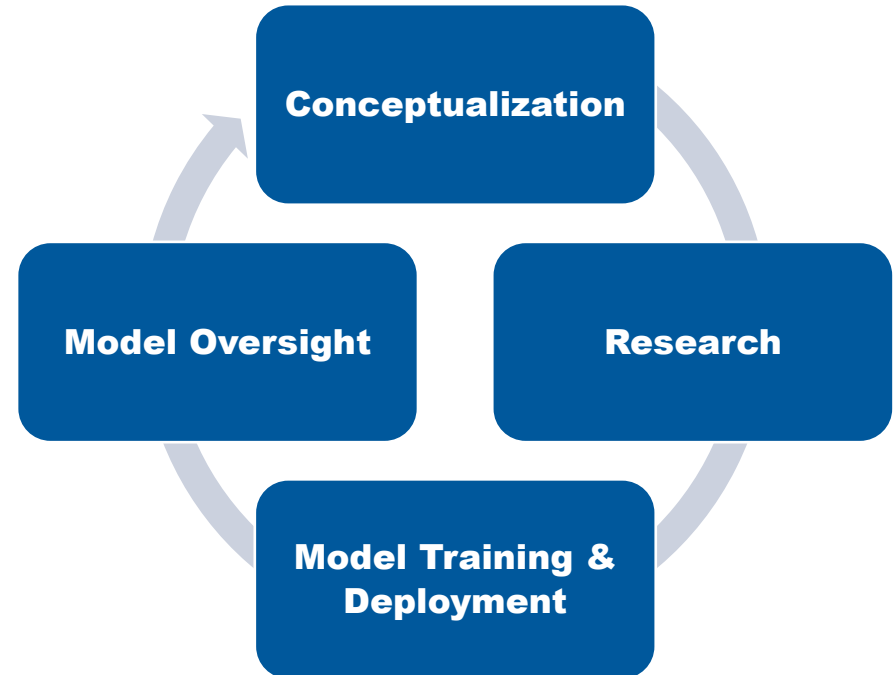
It is in the IRS's best interest that their analysts and management teams are up to date on important principles and practices to promote trust and mitigate risk.

Jameson: Framework for teams

Teams looking to pursue Trustworthy Analytics now have a framework they can follow which:

- Provides **bias mitigation checklists** for each stage of the analytics lifecycle.
- Gives a broad view of how to **design fair systems**.
- Helps teams **affirmatively plan fairness** into systems.

The Analytics Lifecycle





Brian: Toolkit for Analysts

The **analyst toolkit** provides actionable knowledge and steps for analysts to take when assessing fairness and mitigating bias within their models and algorithms.



The goal of the toolkit is to provide analysts with a solid understanding of **points to consider** when assessing and developing models, as well as provide **tools for increasing fairness** in models.



Trustworthy Analytics SharePoint

The image displays a composite of three screenshots from the Trustworthy Analytics SharePoint environment. The top-left screenshot shows the main SharePoint page for 'Trustworthy Analytics', featuring a navigation sidebar with links to Home, Notebook, Documents, Recent, and Site Contents. The main content area includes a 'What is This Resource?' section and a 'Documents' list. The top-right screenshot shows the 'Trustworthy Analytics Notebook' in OneNote Online, with a sidebar listing various sections like 'Introduction Materials', 'Analyst Toolkit', and 'Helpful Tools'. The main content area of the notebook is titled 'Development, Training, and Deployment' and contains definitions for various metrics. The bottom-left screenshot shows a 'Source Material' document titled 'Academic Papers', listing various research papers with their titles, authors, and dates.

Trustworthy Analytics EDIT LINKS

Trustworthy Analytics

Home

Notebook

Documents

Recent

Site Contents

EDIT LINKS

What is This Resource?

This site acts as a repository for efforts underway at the IRS to pursue Trustworthy Analytics, as required by:

1. Informational content describing the principles of Trustworthy Analytics in detail.
2. A Framework assessing the analytics lifecycle for teams looking to insulate their analytic and AI systems.
3. A OneNote which consolidates resources and describes techniques that can be used to detect and prevent bias.
4. A library of literature reviews which informed the resources described above.

Where Do I Start?

If you are not familiar with the concept of Trustworthy Analytics or its guiding principles, start with the Principles of Trustworthy Analytics.

If you are interested in designing

Documents

+ new document or drag files here

Name	Modified
Reviews of Source Material	3 days ago
Source Material	3 days ago

OneNote Online Trustworthy Analytics

File Home Insert Draw View Print Tell me what you want to do Open In OneNote

Find on this Page (Ctrl+F)

+ Page

Initiation and Conceptualization

Research and Design

Development, Training, and Deployment

Fairness Metrics

Preprocessing: Disparate Impact Remover

Preprocessing: Learning Fair Representations

Preprocessing: Optimized Data Preprocessing

Preprocessing: Reweighing

Models - Adversarial Debiasing

Models - Meta-Fair Classifier

Models - Prejudice Remover

Post Process - Equalized Odds

Post Process - Calibrated Equalized Odds

Post Process - Reject Object Classification

Model Card Template

Trustworthy Analytics Notebook

Development, Training, and Deployment

Wednesday, July 13, 2022 11:23 AM

Definitions

Before diving into the technical details of assessing bias and fairness in models, here are some definitions to ground our discussion of Trustworthy Analytics during the Development, Training, and Deployment phase.

Defining Groups

Attribute: $A = \{a_1, a_2, \dots, a_n\}$ (for example, if the attribute is gender, you might see these three values: female, male, other).

Group: $g(a_i)$ is a group of all entities that share the same attribute value (i.e. gender = female)

Reference Group: $g(a_i)$ is one of the groups of A that is used as reference for calculating bias measures

Labeled Positive: L_{P_k} is the number of entities labeled as positive within a group

Labeled Negative: L_{N_k} is the number of entities labeled as negative within a group

Prevalence: $Preval_k = L_{P_k} / |g| = P(Y=1|A=a_i)$, the fraction of entities within a group for which the true outcome was positive.

Distributional Group Metrics

Predicted Positive: PP_k is the number of entities within a group for which the decision is positive

Total Predicted Positive: $K = [Equation]$ is the total number of entities predicted positive across groups defined by A .

Predicted Negative: PN_k is the number of entities within a group for which the decision is negative.

Predicted Prevalence: $PP_{Prev_k} = PP_k / |g|$, the fraction of entities within a group which were predicted as positive.

Predictive Positive Rate: $PPR_k = PP_k / K$, the fraction of the entities predicted as positive that belong to a certain group.

Error Based Group Metrics

False Positive: Number of entities of the group that are falsely labelled as positive

False Negative: Number of entities of the group that are falsely labelled as negative

Source Material EDIT LINKS

Source Material Academic Papers

+ new document or drag files here

All Documents ... Find a file

Name	Modified	Modified By
2022_Stoyanovich_NYU_ResponsibleDataScience_Course	3 days ago	Zheng Chloe (Contractor)
1975_Bickel_SexBiasInGraduateAdminBerkeley	3 days ago	Zheng Chloe (Contractor)
1996_Friedman_Bias in Computer Systems	3 days ago	Zheng Chloe (Contractor)
2012_Kamiran_Decision_Theory_for_Discrimination-Aware_Classification	3 days ago	Zheng Chloe (Contractor)
2016_Friedler_Im-possibilityoffairness-Metrics	8 minutes ago	Zheng Chloe (Contractor)
2016_Hardt_EqualizedOdds-EqualOpportunity	9 minutes ago	Zheng Chloe (Contractor)
2017_Chouldechova_FairPred_Recidivism_Instruments	3 days ago	Zheng Chloe (Contractor)
2017_Pleiss_CalibratedEqualizedOdds	9 minutes ago	Zheng Chloe (Contractor)
2018_Lofus_CausalReasoningForFairness	3 days ago	Zheng Chloe (Contractor)
2018_Selbst_The Inhibitive Appeal of Explainable Machines	3 days ago	Zheng Chloe (Contractor)



Next Steps

- **Literature Review**
 - Refine SharePoint Site and existing guidelines and toolkits
 - Continue to collect and organize resources – expand on topics that were not covered
 - Utilize resources to assist Training Workstream
- **Toolkit Pilot**
 - Assess existing metrics and algorithms to understand best use cases
 - Develop tools training
 - Full implementation and integration of relevant and necessary tools for RAAS Analysts