A Holistic Approach to Trustworthy Artificial Intelligence

Kathryn Matto

Partner, Data and
Technology Transformation

Qingying (Ally) Lu

AI and Business
Transformation Leader

Tara Chavda

Senior Data Scientist

Agenda

- ➢ Challenges of Adopting Trustworthy AI
- ➢ Holistic Approach to AI Governance
- ➢ Assessing AI Models

Challenges in Adopting AI

Organizations are facing a growing set of challenges associated with investments in AI.

- **People:** User Trust and Adoption
- **Process:** Alignment with an evolving regulatory landscape
- **Technology:** AI Behavioral Vulnerabilities

“Earning trust in AI is not a technical challenge but one that is socio-technical and it demands … boldness required by leadership to use the AI as a mirror to their own biases, recognize the calcified systemic inequities in an organization and insist on change.”

Phaedra Boinodiris
IBM hosts AI Equity Summits
IBM **Consulting**

Evolving Regulatory Landscape for Trustworthy AI

Process: Alignment with an evolving regulatory landscape

White House AI Executive Order

Guide for society to protect all people from threats of automated systems that have potential to make an impact to society

NIST AI Risk Management Framework

The framework is to equip organizations and individuals with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, and use of AI systems over time.

AI presents risks which need to be considered & mitigated to scale solutions responsibly and that engenders users' trust

Ethical Considerations in AI

- Bias & Discrimination
- Privacy & Confidentiality
- Transparency & Accountability
- Fair Use
- Erosion of Trust - Hallucinations
- Model Size & Resource Constraints

Responsible AI Models

- Controllable Text Generation
    - Trusted Data Acquisition
    - Domain Adaptation
    - Prompt Engineering
    - Bias Mitigation
    - Adversarial Training
    - Human-in-the-loop
    - Post-Generation Validation
- Transparency & Explainability for AI
- AI Model Impact Assessment

Principles for Trust and Transparency

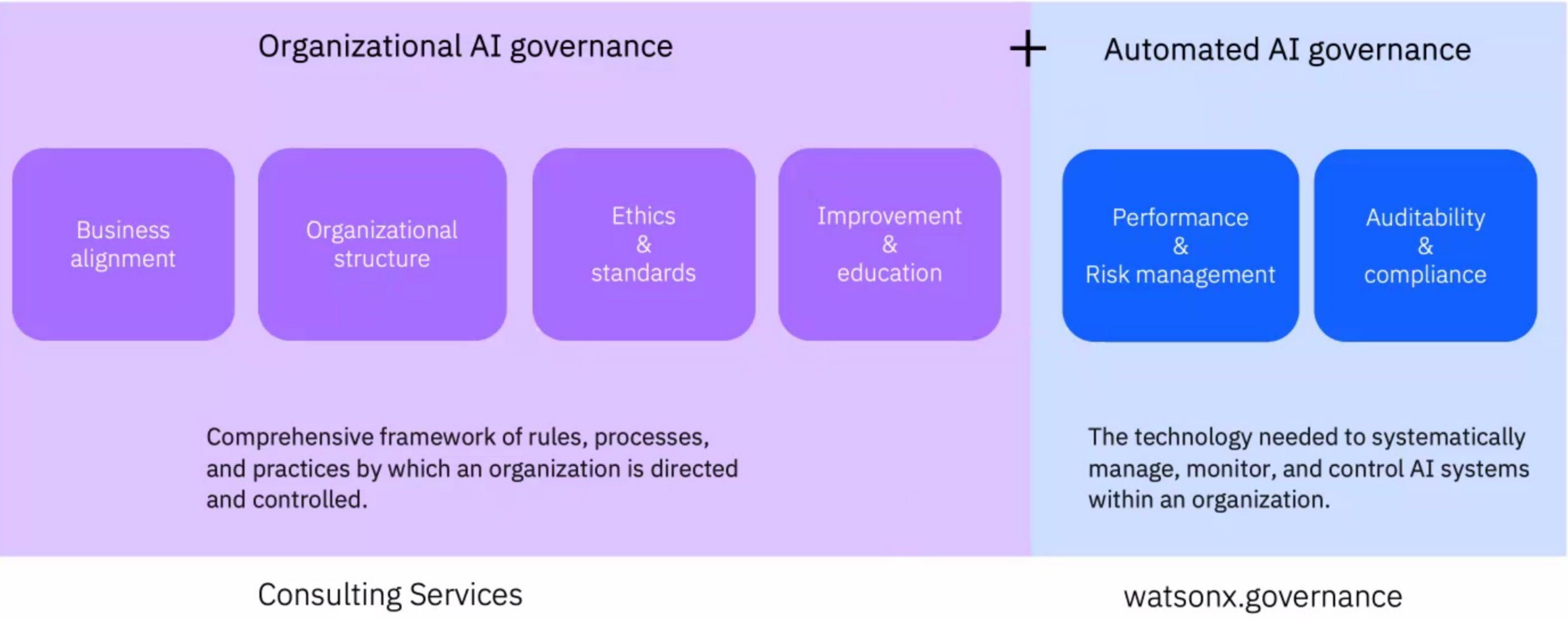
Proposed by IBM

1 The purpose of AI is augment — not replace — human intelligence

2 Data and insights belong to their creator

3 New technology, including AI systems, must be transparent and explainable

Comprehensive Approach to AI Governance



Organizational AI governance + Automated AI governance

Business alignment

Organizational structure

Ethics & standards

Improvement & education

Performance & Risk management

Auditability & compliance

Comprehensive framework of rules, processes, and practices by which an organization is directed and controlled.

The technology needed to systematically manage, monitor, and control AI systems within an organization.

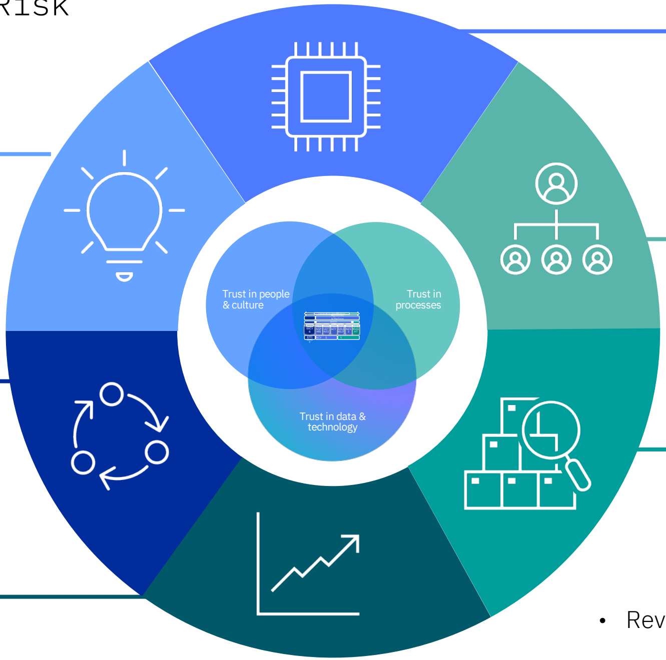
Consulting Services

watsonx.governance

Source: IBM IBV CEO Action Guide: Responsible AI and Ethics

Framework for Responsible AI

Holistic Approach to Mitigate Risk



Awareness & Training

- Understanding unintended consequences
- Engaging and aligning the right stakeholders
- Education for all levels of the business
- Co-develop a comms strategy
- Power Mapping

Operating Model & Adoption

- Design with Ethics and Regulation in mind
- Develop & execute playbook for Trustworthy AI
- Guidance on implementation of AI emerging risks, and safeguards against them

Scaled Practices

- Scaled Data Science practices across the AI lifecycle
- Ensure scalability and smooth operations. Maintain the business benefits of your deployed models.
- Reduce manual efforts to deploy and track model performance. Increase team efficiency
- Feedback loops and Factsheets for transparency and human agency

Principles, Policies & Processes

- Align to business strategy & worldview
- Establish AI Principles & Pillars aligned to values
- Define organizational policies and processes
- Advise on the growing regulatory environment and industry best practices

Organizational structure

- Centralized vs. Federated AI development?
- Define roles & responsibilities of personnel
- Incentive structures
- Establish AI Ethics boards
- Build Advocacy Network and Communities

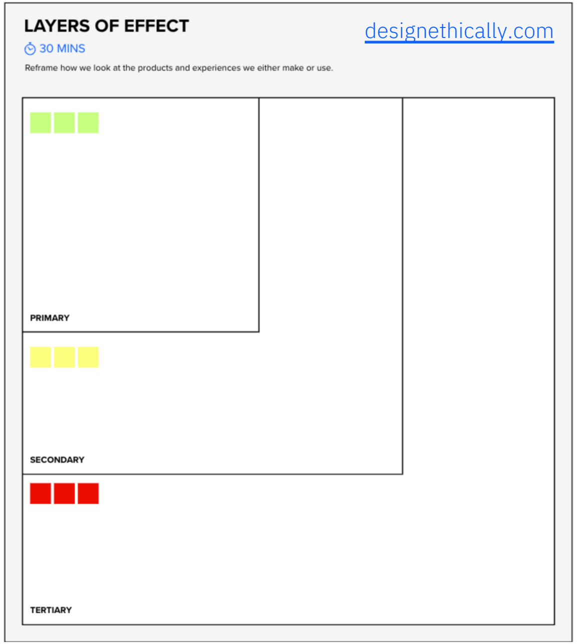
Assess & Audit

- The cost of a Generative AI hallucination
- Risk assessment (initially & regularly) and mitigation
- Consider current & upcoming Data & AI regulation
- Review models, practices, processes, structures, skills and culture for Trustworthiness
- Support and automate internal audits on input and output

IBM

Ethics by Design

How do we look beyond the primary purpose of our AI solution to forecast its effects?



LAYERS OF EFFECT

designethically.com

30 MINS

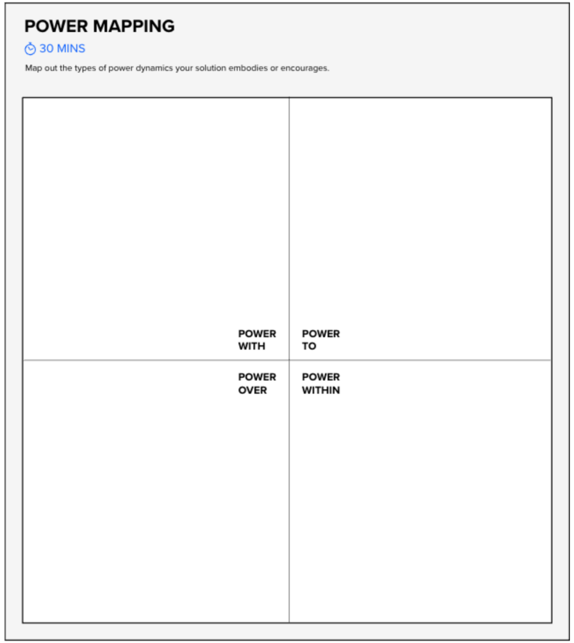
Reframe how we look at the products and experiences we either make or use.

PRIMARY

SECONDARY

TERTIARY

Map out the types of power dynamics your solution embodies or encourages.



POWER MAPPING

30 MINS

Map out the types of power dynamics your solution embodies or encourages.

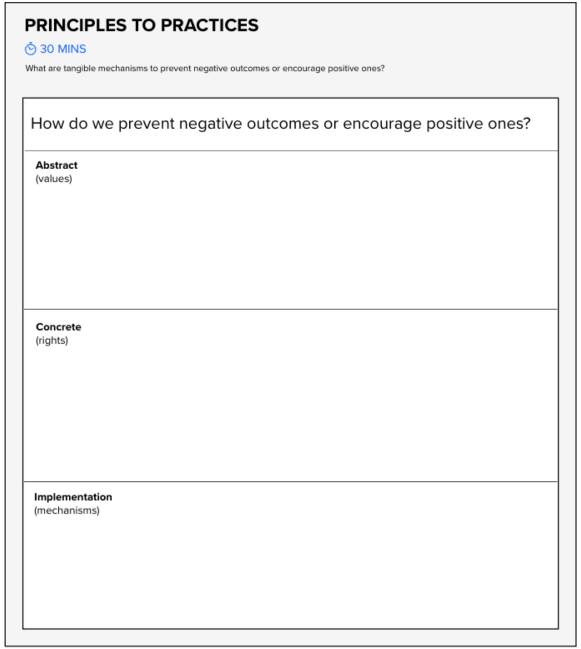
POWER WITH

POWER TO

POWER OVER

POWER WITHIN

What are tangible mechanisms to prevent negative outcomes?



PRINCIPLES TO PRACTICES

30 MINS

What are tangible mechanisms to prevent negative outcomes or encourage positive ones?

How do we prevent negative outcomes or encourage positive ones?

Abstract
(values)

Concrete
(rights)

Implementation
(mechanisms)

Assessing AI Models

- The evaluation of AI models has traditionally focused on gauging their predictive power.

- Instances of unintended consequences have prompted a recognition that there are other crucial dimensions to consider.

*A study published in Science in October 2019 has found that an algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care heavily favored white patients over black patients.**

*The datasets the state-of-the-art dermatology AI models are trained on are limited due to the lack of images of uncommon diseases, as well as diverse skin tones, resulting in their substantial limitations on dark skin tones and uncommon diseases.***

*Ziad Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations. Science 366,447-453(2019).
**Roxana Daneshjou et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci. Adv.8, eabq6147(2022).

A Holistic Approach to Assessing AI models

Competence

Basic performance, i.e., accuracy of an AI model, including uncertainties in the predictions.

Fairness

Equitable treatment of individuals or groups by an AI system

Robustness

An AI systems' ability to maintain good and correct performance across varying operating conditions.

Openness

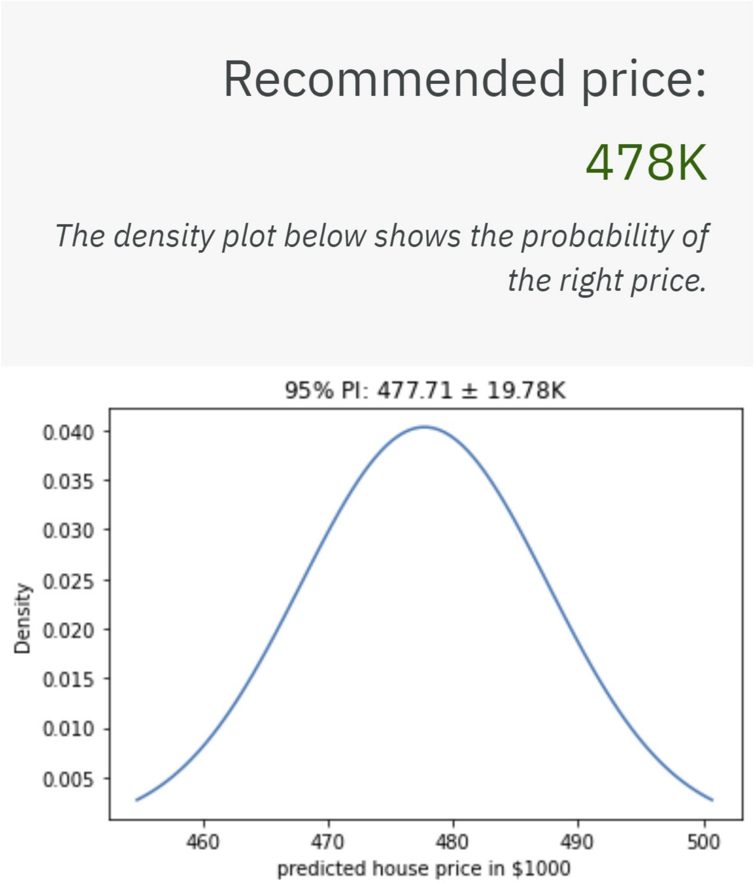
An AI system's ability to include and share information on how it has been designed and developed

Privacy

An AI system's ability to prioritize and safeguard consumers' privacy and data rights

Competence

- The basic performance of an AI/ML model, such as the accuracy of the model
- Quantifying uncertainty around the test results is as important
- Need to test for accuracy under distribution shifts
- IBM Uncertainty Quantification 360 can estimate and evaluate the uncertainty in the AI/ML models

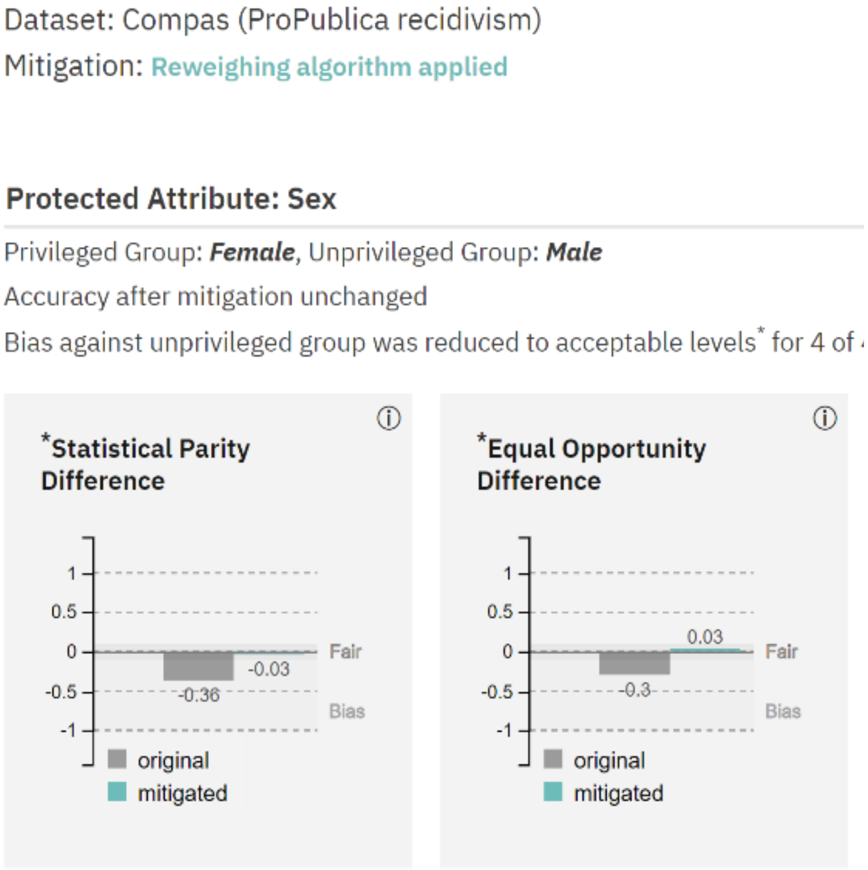


Recommended price:

478K

The density plot below shows the probability of the right price.

Fairness



Dataset: Compas (ProPublica recidivism)

Mitigation: Reweighing algorithm applied

Protected Attribute: Sex

Privileged Group: ***Female***, Unprivileged Group: ***Male***

Accuracy after mitigation unchanged

Bias against unprivileged group was reduced to acceptable levels* for 4 of

- Fairness can be measured at different points in a machine learning pipeline
  - On the training data
  - On the learned model, which also relates to the pre-processing, in-processing, and post-processing categories of bias mitigation algorithms
- IBM AI Fairness 360 can examine and evaluate discrimination and bias in AI/ML models, including age bias, racial bias, etc., and provide mitigation recommendations

Reliability

- Includes safety and security of AI/ML models and systems.
- AI/ML systems need to maintain good and correct performance across varying operating conditions
- Adversarial attacks refer to crafted alterations included in the input that leads an AI/ML model to make incorrect predictions
- IBM Adversarial Robustness 360 can defend and verify AI/ML models against adversarial attacks

Openness

- Applies to various aspects related to human interaction with AI/ML models
  - Communication from the machine to the human through interpretability and explainability of models by people
  - System reporting its quantified uncertainty or confidence
  - Transparency into overall AI/ML system pipelines and lifecycles
- IBM AI Explainability 360 includes many different algorithms to explain data vs. model, directly interpretable vs. post hoc explanation, local vs. global, static vs. interactive
- AI FactSheets captures model or service facts from the entire AI lifecycle and are compiled with inputs from multiple roles in this lifecycle.
- IBM Design Guide offers guidance on how to best write AI FactSheets

Privacy

- A malicious party with access to a trained ML model can still reveal sensitive, personal information in the training data even without access to the training data itself.
- It is therefore crucial to be able to recognize and protect AI models that may contain personal information.
- AI Privacy 360 can assess and protect against the privacy risks of the AI/ML models. The tool can also provide recommendations for adhering to any relevant privacy requirements by exploring tradeoffs between privacy, accuracy, and performance of the model

- Random Forest model predicting mortality of hepatitis patients
- The model accuracy was 0.8 and the attack accuracy was 0.56, determined by AI Privacy 360, which means that some membership information was leaked (the attack accuracy exceeds randomly guessed predictions of 0.5).
- After model anonymization is applied in AI Privacy 360, the attack accuracy dropped to 0.48 which means the privacy leakage was addressed while the model accuracy stays at a similar level,

	Model Accuracy	**Attack Accuracy**
Original	0.8	0.56
Mitigated	0.79	0.48

Open-Source Trustworthy AI Toolkits

Uncertainty Quantification 360

Comprehensive open-source toolkit for computing and communicating meaningful limitations of ML predictions.

AI Fairness 360

Comprehensive open-source toolkit to help detect & mitigate bias in ML models.

Adversarial Robustness 360

Comprehensive open-source toolkit for defending AI from attacks.

AI Explainability 360

Comprehensive open-source toolkit for explaining ML models & data.

AI Factsheets 360

Extensive website describing research efforts to foster trust in AI by increasing transparency and enabling governance.

AI Privacy 360

Toolbox to support the assessment of privacy risks of AI-based solutions, and to help them adhere to relevant privacy requirements.

The platform for Trusted AI and Data

watsonx

Scale and accelerate the impact of AI with trusted data.

watsonx.ai

Train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store optimized for governed data and AI workloads, supported by querying, governance and open data formats to access and share data.

watsonx.governance

Enable responsible, transparent and explainable data and AI workflows

End-to-end toolkit encompassing both data and AI governance to enable responsible, transparent, and explainable AI workflows.

Conclusion

Every person involved in the creation of AI at any step is accountable for considering the system’s impact in the world.

Curating AI is a socio-technical challenge that, to solve, requires a **holistic approach** encompassing people, processes and tools.

Assessing AI models multi-facet to **build trust** throughout the lifecycle.

Thank you!

Kathryn Matto
kathryn.matto@us.ibm.com

Qingying (Ally) Lu
qingying.lu@us.ibm.com

Tara Chavda
tara.chavda@ibm.com

IBM **Consulting**