RESPONSIBLE AI

OPPORTUNITIES AND CHALLENGES FOR TESTERS

Bill Matthews

AI RISKS AND HARMS

2023 survey of 800 businesses and IT decision makers found that

85%

Expect AI to increase revenue growth in the next 18-24 months

36%

Were confident their organization has sufficient checks and balances in place to mitigate potential risks and harms of AI

73%

Agreed that Safe AI and Responsible AI are a priority for their organization over the next 18-24 months

Source https://www.avanade.com/-/media/asset/thinking/avanade-trendlines-exec-summary-pdf.pdf

WHAT IS RESPONSIBLE AI

RESPONSIBLE AI SEEKS TO...

- Provide a set common themes that aim to promote AI that is Useful and Beneficial to Society.
- Mitigate risks and harms of AI
- Foster a more humanistic approach to building AI systems
- Build Trust in AI Systems
- Supports Regulatory Compliance

RESPONSIBLE AI IS NOT...

- Solely about meeting regulatory and governance requirements
- Solely about Ethics
- A collection of best practices for companies to adopt

It is multi-disciplinary, context-driven, risk focused and people centric

COMMON THEMES IN RESPONSIBLE AI

Fairness

- Built for Everyone
- Avoid creating or reinforcing Bias and Discrimination
- Fairness of outcomes

Transparency and Explainability

- Ensuring how the AI System is used is understood.
- Ensuring AI Decisions can be Explained and Understood

Accountability

- Clear responsibility for AI Actions and Decisions
- People are accountable not algorithms

Privacy and Security

- Respects user data and confidentiality
- Avoids data leakage

Robustness and Reliability

- Reliable outcomes
- Safe outcomes
- Robust to malicious users

THE IMPORTANCE OF RESPONSIBLE AI

WHY IT MATTERS

- Impacts on individuals and society

REAL WORLD CONSEQUENCES

British female politicians targeted by fake pornography

Leading politicians victimised by online material including AI deepfakes, investigation finds

Victims include Labour's deputy leader, Angela Rayner (left) and the Commons leader, Penny Mordaunt (right). Photograph: Getty Images

British female politicians have become the victims of fake pornography, with some of their faces used in nude images created using artificial intelligence.

https://www.theguardian.com/technology/article/2024/jul/01/british-female-politicians-targeted-by-fake-pornography

THE IMPORTANCE OF RESPONSIBLE AI

WHY IT MATTERS

- Impacts on individuals and society
- **Business Reputation**

REAL WORLD CONSEQUENCES

Air Canada Has to Honor a Refund Policy Its Chatbot Made Up

The airline tried to argue that it shouldn't be liable for anything its chatbot says.

PHOTOGRAPH: ROBERT SMITH/GETTY IMAGES

https://www.wired.com/story/air-canada-chatbot-refund-policy/

THE IMPORTANCE OF RESPONSIBLE AI

WHY IT MATTERS

- Impacts on individuals and society
- Business Reputation
- **Regulatory Compliance**

REAL WORLD CONSEQUENCES

Artificial Intelligence Act: MEPs adopt landmark law

Press Releases PLENARY SESSION IMCO LIBE 13-03-2024 - 12:25

- Safeguards on general purpose artificial intelligence
- Limits on the use of biometric identification systems by law enforcement
- Bans on social scoring and AI used to manipulate or exploit user vulnerabilities
- Right of consumers to launch complaints and receive meaningful explanations

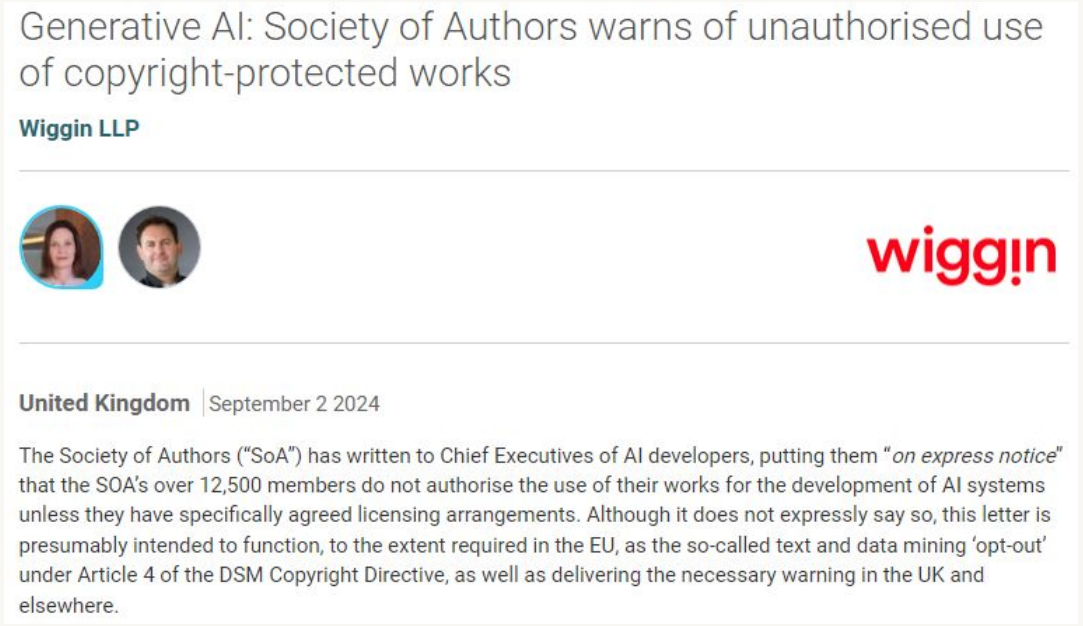
The untargeted scraping of facial images from CCTV footage to create facial recognition databases will be banned © Alexander / Adobe Stock

THE IMPORTANCE OF RESPONSIBLE AI

WHY IT MATTERS

- Impacts on individuals and society
- Business Reputation
- Regulatory Compliance
- **Legal Challenges**

REAL WORLD CONSEQUENCES



Generative AI: Society of Authors warns of unauthorised use of copyright-protected works

Wiggin LLP

wiggin

United Kingdom | September 2 2024

The Society of Authors ("SoA") has written to Chief Executives of AI developers, putting them "*on express notice*" that the SOA's over 12,500 members do not authorise the use of their works for the development of AI systems unless they have specifically agreed licensing arrangements. Although it does not expressly say so, this letter is presumably intended to function, to the extent required in the EU, as the so-called text and data mining 'opt-out' under Article 4 of the DSM Copyright Directive, as well as delivering the necessary warning in the UK and elsewhere.

https://www.lexology.com/library/detail.aspx?g=65ce4bc7-bea2-4706-a6fd-f589776b6265

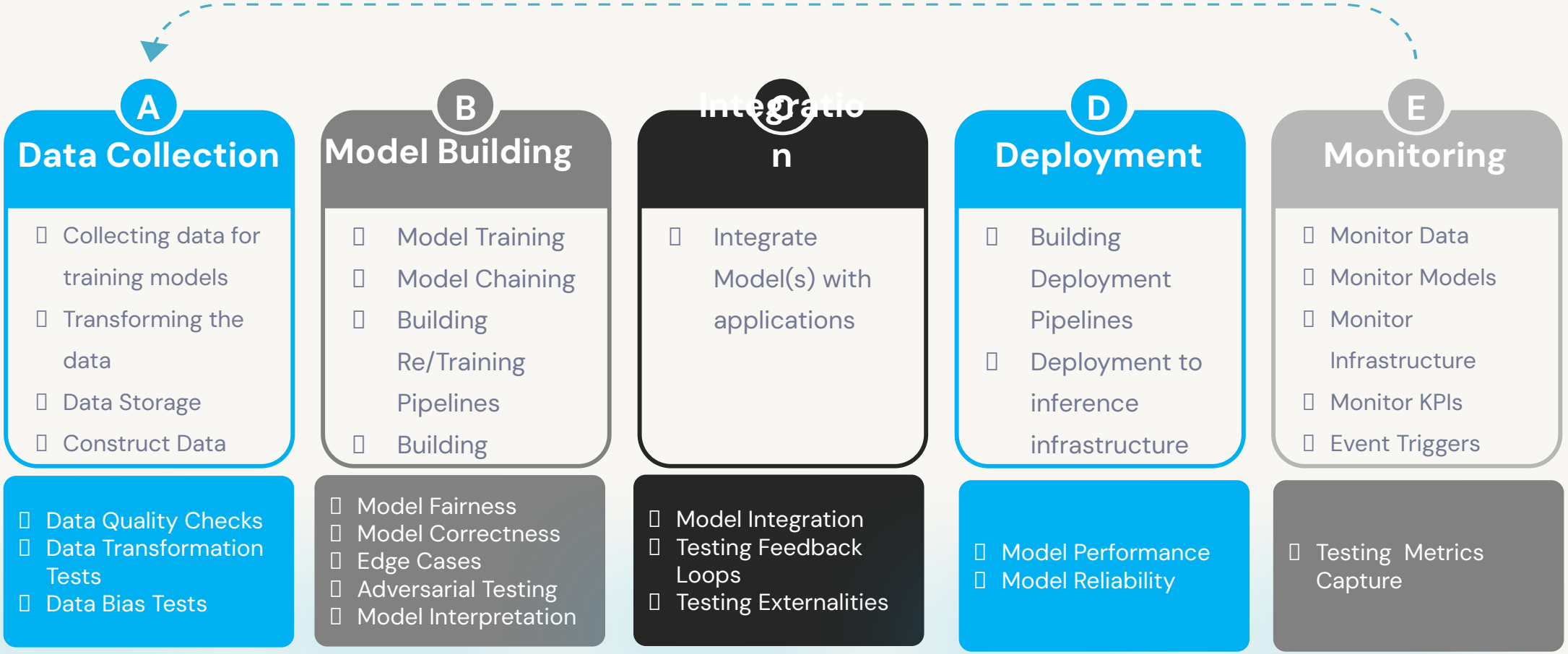
WHAT DOES THIS HAVE TO WITH TESTING?

Developing Responsible AI requires an understanding of the possible issues, limitations, and unintended consequences of the AI System

Sounds a lot like testing to me

THE ROLE OF TESTING IN RESPONSIBLE AI

WHERE (SHOULD) TESTING OCCUR TO SUPPORT RESPONSIBLE AI?



WHERE (MIGHT) TESTERS FIT INTO RESPONSIBLE AI?

MODEL TESTING

- Model Fairness and Bias
- Model Correctness
- Model Reliability
- Adversarial Testing
- Edge Case Testing
- Privacy and Security Tests

SYSTEM TESTING

- Positive Feedback Loops
- Negative Feedback Loops
- Data Flows

CHALLENGES FOR TESTERS

KNOWLEDGE

- Foundational AI/ML
- Responsible AI Themes
- AI/ML specific Technical Risks and manifestations
- AI/ML specific testing techniques
- Data Analysis and manipulation

MINDSET

- Notions of Correctness
- Testing distributions not data points
- Hypothesis Testing

TOOLING

- Tooling tends to be Code-First tools
- Some No/Low Code tools are emerging
- Tool space is fast evolving

TESTING FOR FAIRNESS AND BIAS

ENSURING EQUITABLE OUTCOMES FOR DIFFERENT GROUPS

General Approach

- Identify the different groups that are important in your context
- Understand how membership of those groups manifest in your input data
- Understand the ways that Bias and Un-Fairness can manifest in your model output (threats to Fairness)
- Understand potential harms for specific groups and how that might manifest in your models
- Choose how you will assess fairness and bias

Common Approaches to Fairness and Bias Testing

Group Fairness - ensuring the model's performance is consistent across different demographic groups.

- **Demographic Parity:** Measures whether different groups receive positive outcomes at equal rates
- **Equal Opportunity:** Evaluates whether a model's true positive rate (sensitivity) is the same for all groups.
- **Equalized Odds:** Extends equal opportunity by ensuring that both the true positive rate and false positive rate are equal across groups.

Individual Fairness - focuses on treating similar individuals similarly

- **Counterfactual Fairness:** Measures whether a model's decision would remain the same if an individual's group membership were changed while keeping all other attributes constant
- **Similarity Metrics:** Assess whether similar individuals (based on specific feature similarities) receive similar outcomes

Causal Fairness - examines the causal relationships between inputs to ensure that protected attributes do not unjustly influence outcomes.

- **Total Effect:** Measures the total influence of a protected attribute on the model's prediction
- **Direct and Indirect Effects:** Decomposes the influence of a protected attribute into direct and indirect effects to identify and mitigate unfair causal pathways

TESTING FOR CORRECTNESS

ENSURING THE RELIABILITY OF THE MODEL

General Approach

- Determine the properties of correctness in your context
- Determine how correctness will be measured
- Determine how the inputs will be generated
- Determine how the inputs and outputs could be segmented into useful segments
- Understand what are the expectations based on domain knowledge

Accuracy/Precision/Recall – Assess how well the model performs against the data (often evaluated as part of model training)
Segmentation Analysis – Assess how well the model performs against different segments of inputs/outputs.
Error Analysis – Investigating the cases where the model produced incorrect or unsatisfactory outputs.
Expectation Testing – Does the model hold true with respect to the identified business expectations.

TESTING FOR MODEL ROBUSTNESS

BUILDING TRUST IN THE MODEL

- **Adversarial Testing -** intentionally exposing the model to adversarial inputs—slightly modified examples designed to deceive the model or to control the outcome.
  - **Attack Success Rate**: Measures the effectiveness of adversarial attacks against the model.
  - **Robust Accuracy**: Assesses model performance on adversarial perturbed inputs compared to normal inputs
- **Noise Injection/Perturbation/Augmentation Testing –** intentionally injecting small changes/noise into inputs to understand how they impact the outcomes
  - **Performance Degradation**: Measures the drop in accuracy, precision, or other performance metrics when noise is added to the inputs
  - **Noise Sensitivity**: Quantifies how much input noise affects the model's predictions
- **Out-Of-Distribution (OOD) Testing –** Assess the model's performance using data that is outside the norms of the training distribution.
  - **OOD Detection Rate**: Measures the model's ability to correctly identify when an input is out-of-distribution.
  - **Confidence Calibration**: Evaluates whether the model's confidence in its predictions appropriately reflects its uncertainty on OOD inputs.

TESTING FOR EDGE CASES

TESTING THE BOUNDARIES

General Approach

- Identify the edge cases (that matter) for your context
- Understand how these are represented in the input data
- Understand the potential impact of failures

- **Edge Case Testing –** Assess the model's performance using data that is outside the norms of the training distribution.
  - **Edge Case Detection Rate**: Measures the model's ability to correctly handle an Edge Case.
  - **Confidence Calibration**: Evaluates whether the model's confidence in its prediction of an Edge Case is comparable to typical cases.

TESTING FOR FEEDBACK LOOPS

ENSURING MODEL STABILITY OVER TIME

General Approach

- Identify the feedback loops within interactions
- Identify the positive feedback loops and recognising how these can be detected in the output
- Identify the negative feedback loops and recognising how these can be detected in the output

Feedback Loop Testing - simulating feedback loops to understand how model outputs can affect future inputs. Sensitivity analysis examines how variations in input can lead to changes in model performance

- **Sensitivity Metrics:** Measures how sensitive model outputs are to changes in inputs, particularly those generated by the model itself.
- **Error Propagation Analysis:** Evaluates how errors can accumulate through feedback loops, affecting long-term model quality.

Counterfactual Testing- involves altering specific inputs to observe how these changes affect model outputs, particularly focusing on inputs that are influenced by the model itself in feedback scenarios.

- **Counterfactual Impact Score:** Measures the influence of altered inputs on model predictions.
- **Bias Amplification Metrics:** Evaluates whether feedback loops cause biases to grow over time.

TESTING FOR DATA FLOWS

ENSURING LONGEVITY OF MODELS

General Approach

- Understand where and how data is collected for training future versions of models

Common Approaches to testing ML Data Flows

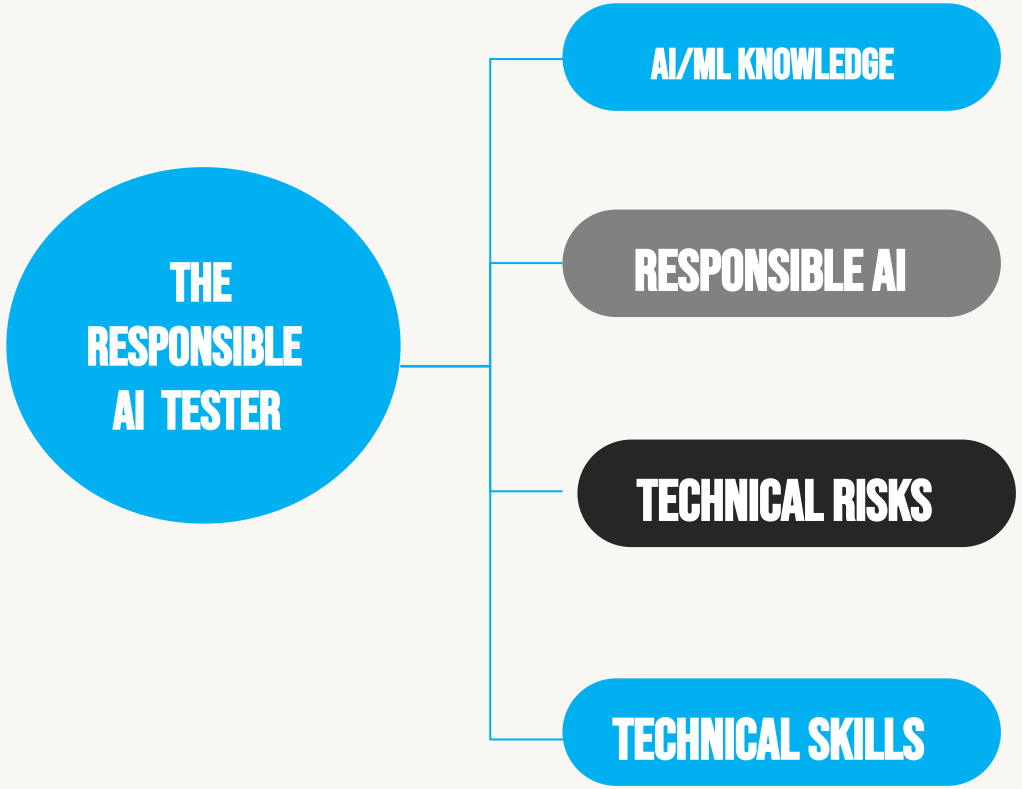
Data Poisoning- Simulate new data arriving from source that aims to artificially change the distribution of the data (such as using outliers, toxic inputs etc.)

- **Population Stability Index (PSI):** Measures the stability of feature distributions over time.
- **Kullback-Leibler Divergence:** Quantifies the difference between the probability distributions of the training data and the current data.

Population Bias – Simulate biased (but within the current distribution) data arriving from source that aims to artificially change the distribution of the data

- **Population Stability Index (PSI):** Measures the stability of feature distributions over time.
- **Kullback-Leibler Divergence:** Quantifies the difference between the probability distributions of the training data and the current data.

CHARTING YOUR ROADMAP INTO RESPONSIBLE AI



- Has foundational knowledge of AI/ML.
- Understands the engineering practices used to build and deploy AI/ML Models

---

- Understanding the RAI themes
- How they apply to products
- How they apply to testing practices

- Understands the **Technical risks** related to using AI in products.
- Understands the ways AI **systems can fail** (manifestation of technical risks)

- Some skills in **Data Analysis, Uncertainty, Hypothesis Testing**
- Familiarity with using common AI Testing tools
- Familiarity with AI Testing Techniques

THANK YOU!

Ask Questions now or later via:

- During TestBash (I'm very approachable!)
- Ministry of Testing Slack
- LinkedIn: **linkedin.com/in/billmatthews**
- Threads: **@BillMatthews**