AI Auditing

Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing

Inioluwa Deborah Raji*
Partnership on AI
deb@partnershiponai.org

Andrew Smart*
Google
andrewsmart@google.com

Rebecca N. White
Google

Margaret Mitchell
Google

Timnit Gebru
Google

Ben Hutchinson
Google

Jamila Smith-Loud
Google

Daniel Theron
Google

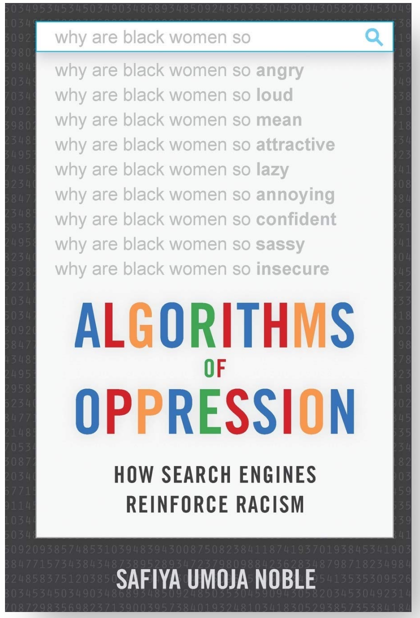
Parker Barnes
Google

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20). Association for Computing Machinery, New York, NY, USA, 33–44.*

What is an audit?

- Audits are tools for interrogating complex processes to determine whether they comply with company policy, industry standards or regulations.

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4



Why Internal Auditing?

- Deployed systems are audited for harm by investigators from outside the organizations.
- For data practitioners, it may be challenging to identify ethically significant consequences.
- The authors introduce a framework for algorithmic auditing that could be used throughout the development life-cycle.
- The goal is to close the accountability gap in the development and deployment of AI systems.

SMACTR: An Internal Audit Framework

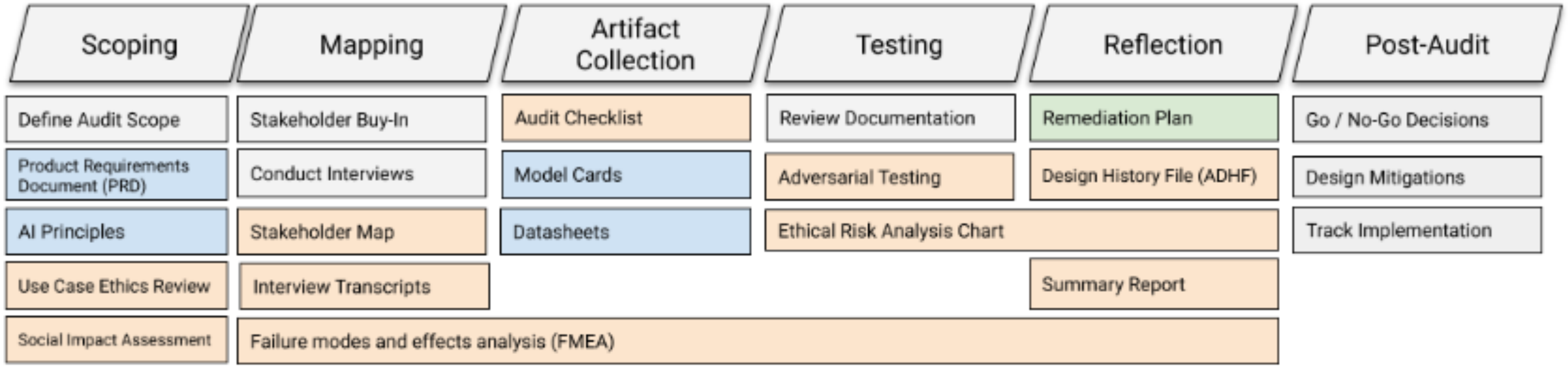
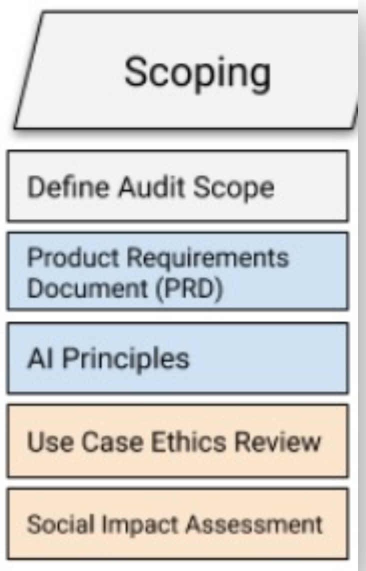


Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

SMACTR: Scoping Stage

- Clarifying the objective of the audit,
- Reviewing the motivations and intended impact of the investigated system,
- Confirming the principles and values meant to guide product development.



SMACTR: Mapping Stage

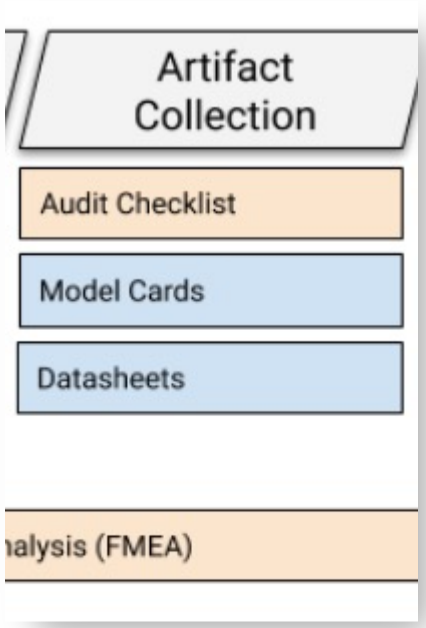
- Checking the perspectives involved in the audited system.
- Failure modes and effects analysis (FMEA) starts in this stage.
- Semi-structured interviews should be conducted with people close to the development process.
- Risks should be prioritized for later testing.

"To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore to make a category error, or as we posit here, an abstraction error."

Selbst, Andrew D. and Boyd, Danah and Friedler, Sorelle and Venkatasubramanian, Suresh and Vertesi, Janet, Fairness and Abstraction in Sociotechnical Systems (August 23, 2018). 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*), 59-68

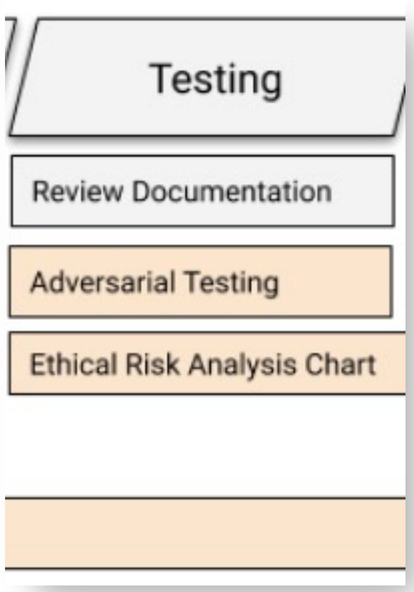
SMACTR: Artifact Collection Stage

- Identifying and collecting all the required documentation from the product development process.
- Documentation can be distributed across different teams and stakeholders.
- The audit checklist is the main artifact in this stage.



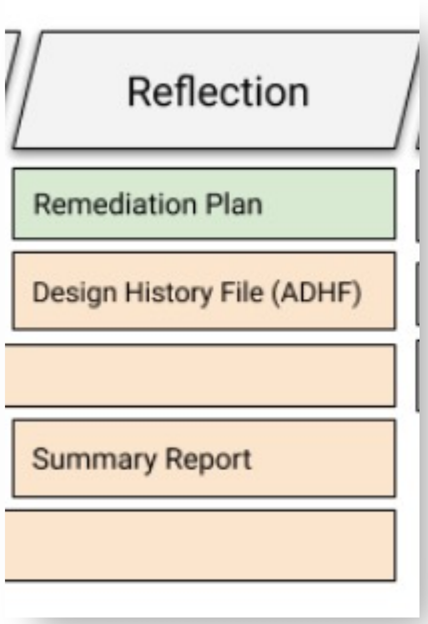
SMACTR: Testing Stage

- The active testing activity starts here.
- Testing is based on a risk prioritization from the FMEA.
- Adversarial testing focuses in finding vulnerabilities.
- Adversarial testing also informs ethical risk analysis to identify the severity of a failure.



SMACTR: Reflection Stage

- Testing results are analyzed considering ethical expectations clarified in the audit scoping.
- The main artifact is a mitigation plan jointly developed by the audit and engineering teams.
- The summary (audit) report should be compared qualitatively and quantitatively to the ethical expectations.



SMACTR: An Internal Audit Framework

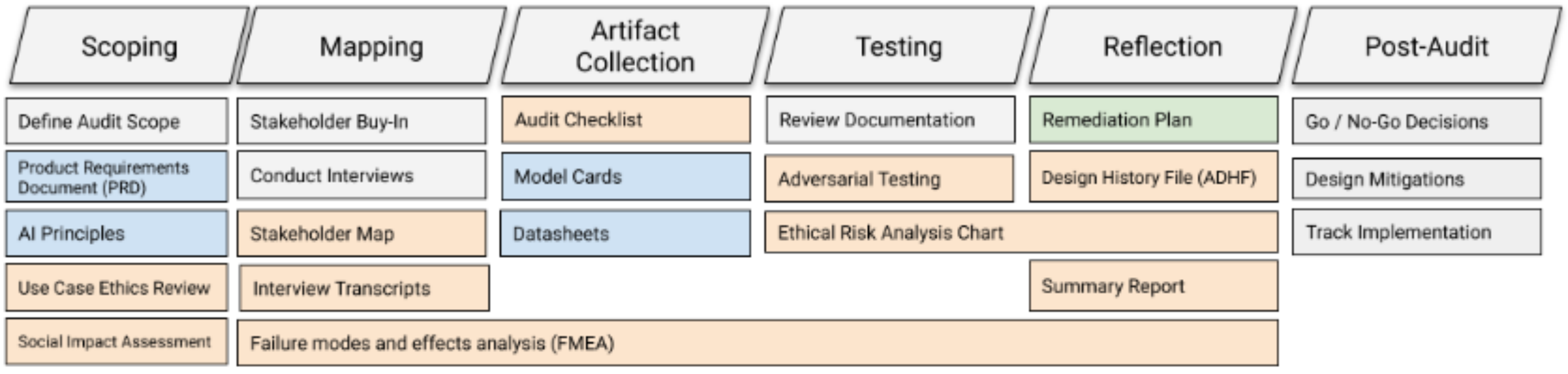


Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

ICO - Guidance on the AI Auditing Framework

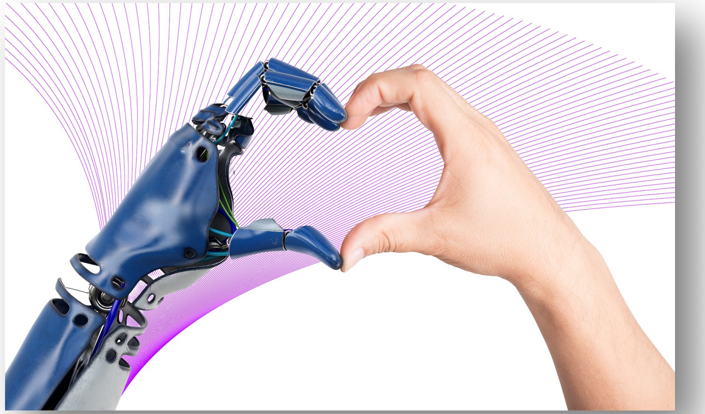
A Risk-based Perspective

Version 1: https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Guidance Outline

- ICO is focusing on a risk-based approach to AI
  - Assessing the risks to the rights and freedoms of individuals that may arise
- Guidance prepared for:
  - an audience with a compliance focus (e.g., data protection officers (DPOs), ICO's own auditors)
  - technology specialists (e.g., developers)
- Four main topics are covered:
  - Accountability and governance of AI
  - Fair, lawful, transparent processing
  - Data minimisation and security
  - Rights in AI systems
- Controls: Preventative, Detective, Corrective

Part I: Accountability and Governance of AI



- Data protection impact assessments (DPIAs)
  - How you will collect, store and use data;
  - The volume, variety, and sensitivity of the data;
  - The nature of your relationship with individuals;
  - The intended outcomes for individuals/society;
  - Data processing steps (what data, the number of data subjects, the source of data, error analysis based on fairness metrics etc.);
  - What could the potential risks be?

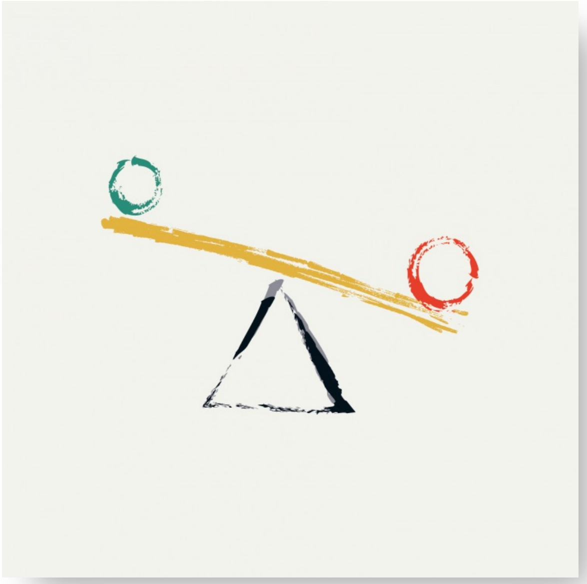
- Senior management, including DPOs, are accountable for understanding and addressing technical complexities of AI systems.

Part I: Accountability and Governance of AI

- Controller/joint controller/processor responsibilities
    - Controller decides on the purposes and means of processing
    - Processor works with personal data under the instruction of another organisation
    - Joint controllers determine the purposes and means of processing with another organisation
- Personal data is processed at several different phases, you may have different roles for some of the phases.

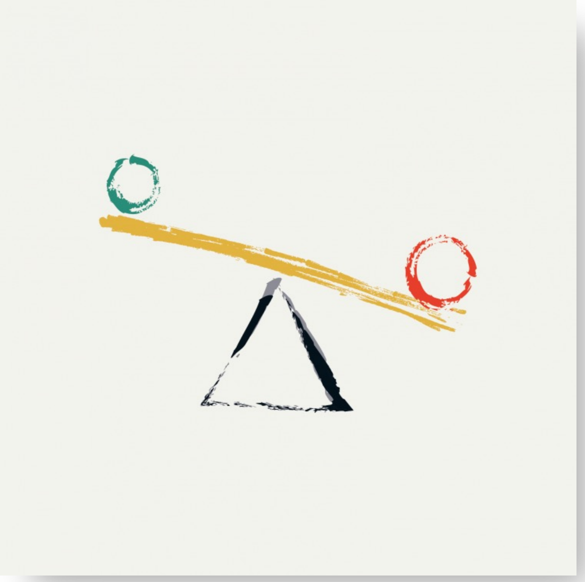
Part I: AI-related trade-offs based on social context

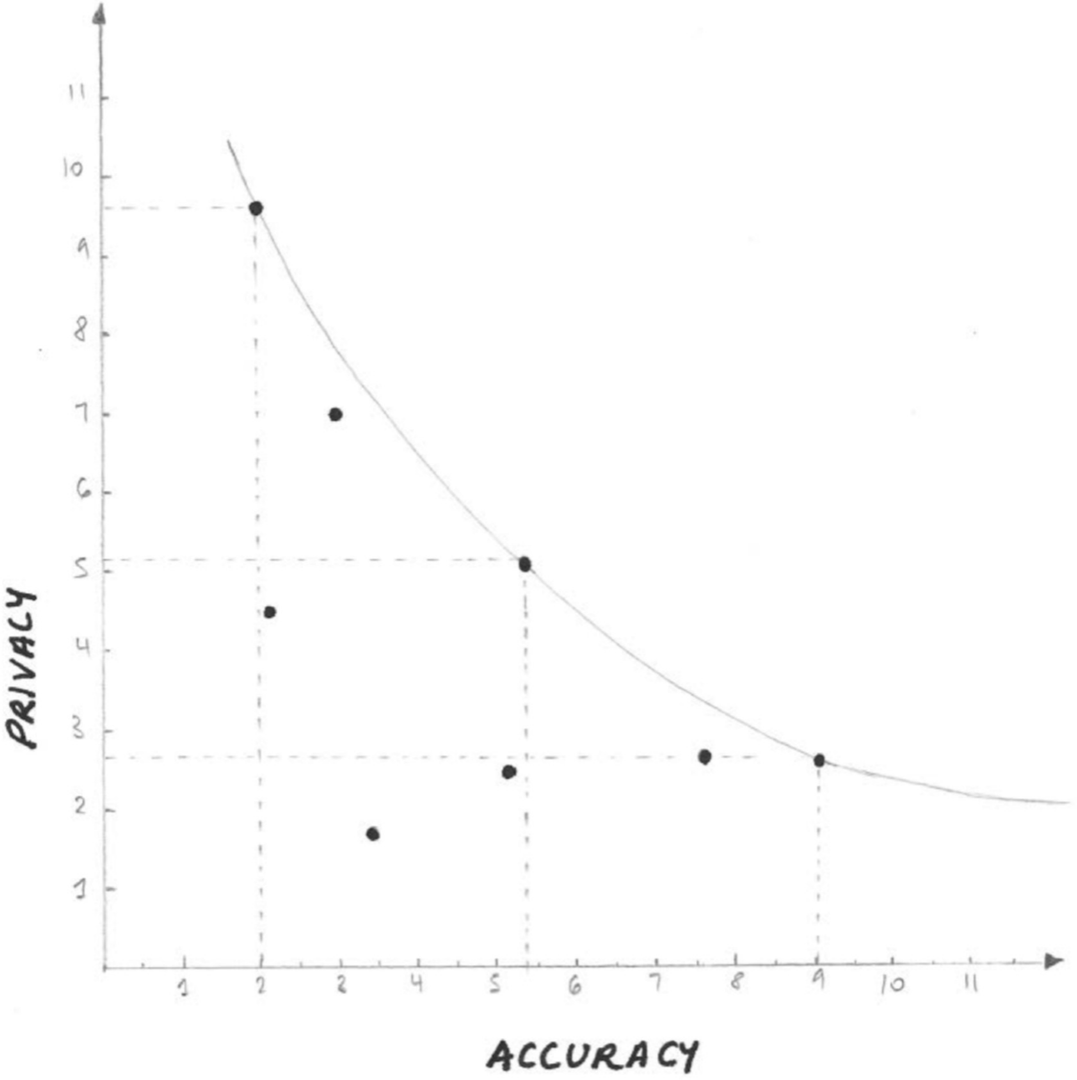
- Privacy vs statistical accuracy
  - Collecting more data points about each person -> greater risks
  - Improving statistical accuracy -> compliance with the fairness principle
- Statistical accuracy and discrimination
  - Preventing discriminatory outcomes -> increasing statistical errors (e.g., statistical parity)

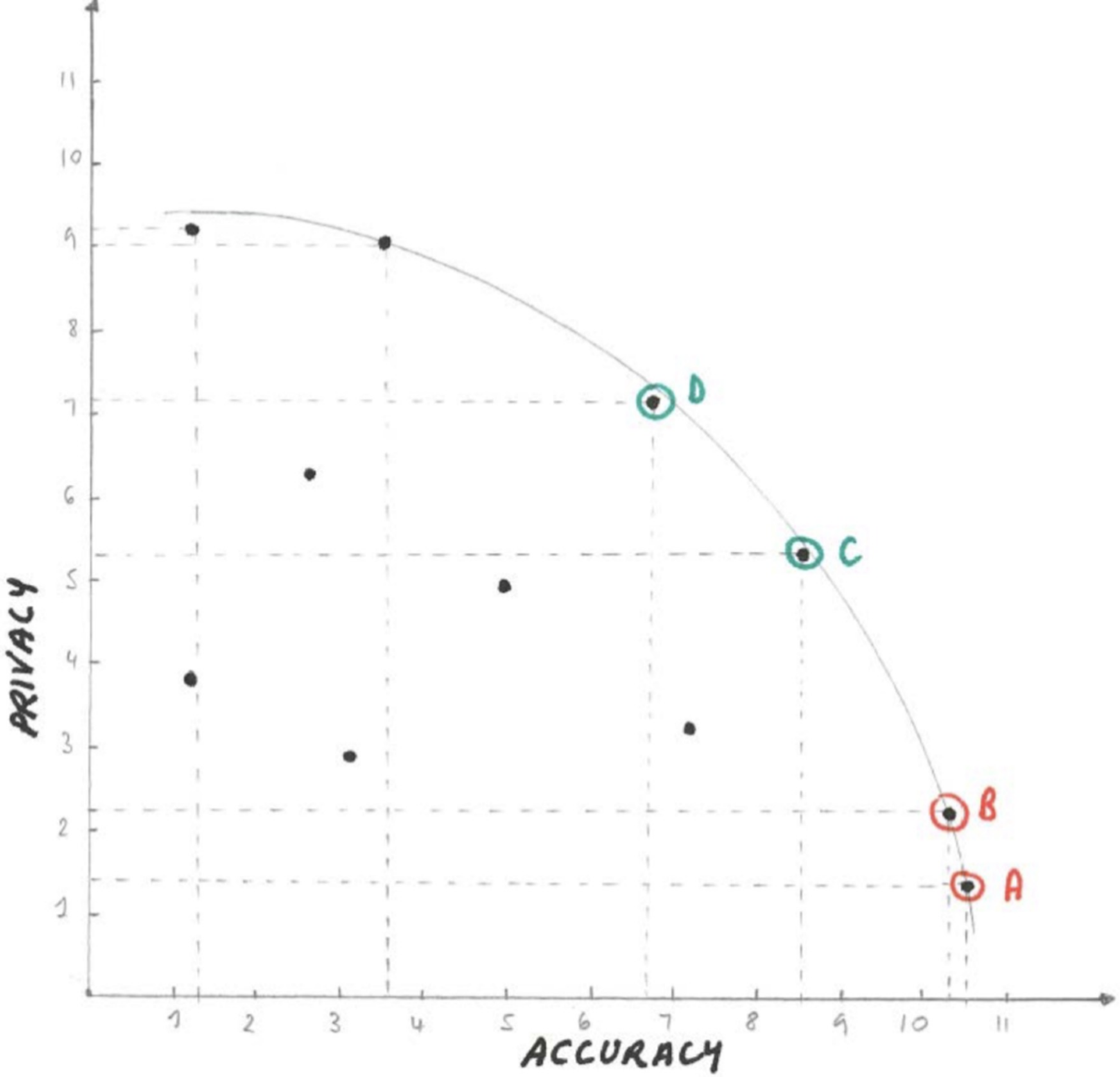


Part I: AI-related trade-offs based on social context

- Explainability and statistical accuracy
  - Black box models, accurate but non-explainable models (e.g., image recognition)
  - ExplAIn project guidance (use black box models if you are aware of the risks, and you have tools to interpret the results with some level of explainability)
- Explainability, exposure of personal data, and commercial security
  - Disclosing personal information while providing explanations (e.g., attacks on trained models)
  - Disclosing proprietary information about how AI works



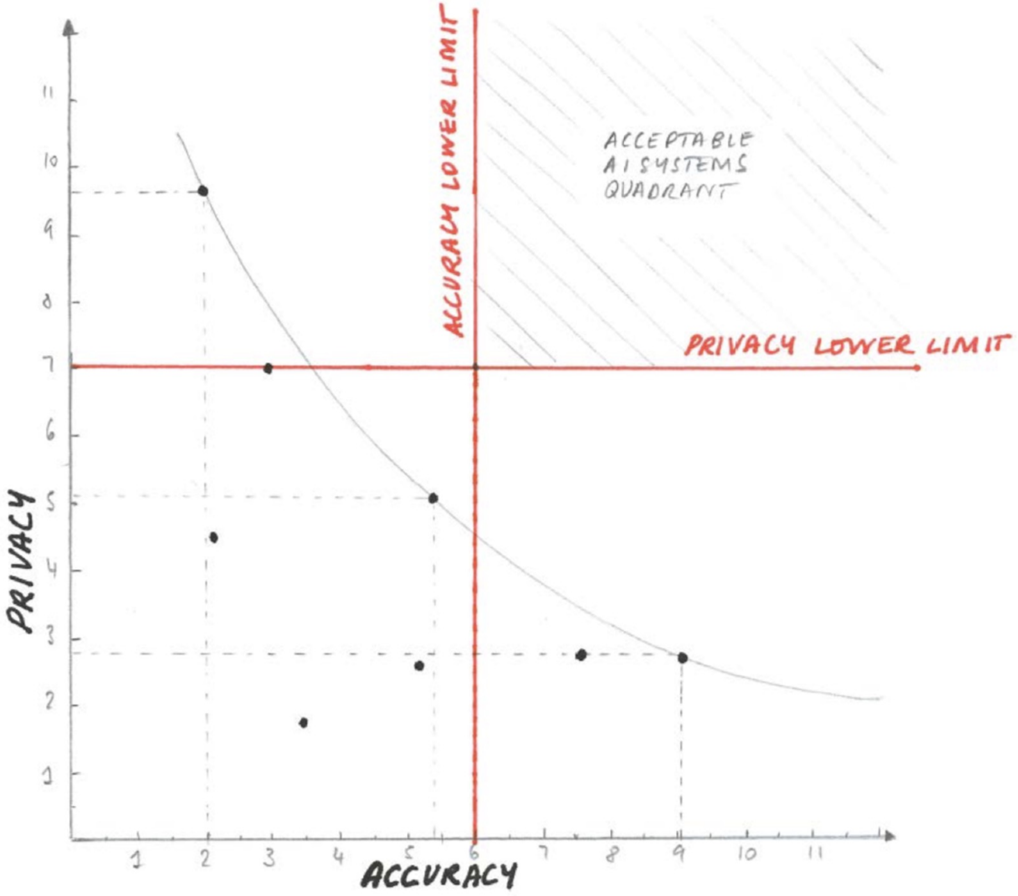




Privacy vs Accuracy

Privacy vs Accuracy

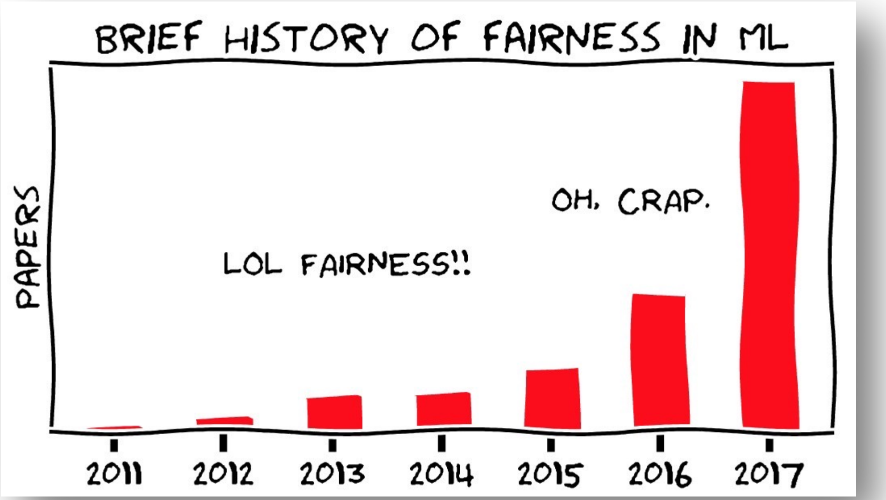
- There is no AI system satisfying lower limits.
- This system should not be deployed.
- What to do?
  - Use other methods/data sources
  - Reformulate the problem
  - Don't attempt to use AI to solve this!



Part II: Fair, lawful, transparent processing

- Lawful bases defined in Article 6 of the GDPR:
  - Consent, contract, legal obligation, vital interests, public task, legitimate interests

- Lawful bases for processing personal data
  - should be decided at the beginning
  - should be included in the privacy notice
  - different for development/deployment phases



Part II: Fair, lawful, transparent processing

Assessing and improving AI system performance

- Statistically informed guesses should be recorded separately
- The provenance of data and AI used to generate the inference should be recorded
- Recording inferences based on inaccurate data is important
- Checking statistical accuracy over time is needed (danger: concept/model drift)

Part II: Fair, lawful, transparent processing

- Mitigating potential discrimination
  - imbalanced training data problem
  - training data reflecting past discrimination
    (danger: proxy variables)

- Fairness measures (not compatible with each other)
  - Anti-classification (excluding protected characteristics)
  - Outcome / error parity
    (equal numbers of positive/negative outcomes; equal numbers of errors to different groups)

Part II: Fair, lawful, transparent processing

Mitigating the risks

- Working with representative data
- Senior management is responsible for signing-off the chosen approach to manage discrimination risk; and be accountable for its compliance with data protection law.
- Robust testing, monitoring, risk management policies/organisational policies should be in place

Part III: Data minimisation and security

- Two security risks:
  - loss or misuse of the large amounts of personal data
  - software vulnerabilities to be introduced

- Data sharing risks (with internal/external entities)
- Security risks introduced by externally maintained software

Part III: Data minimisation and security

Mitigating the risks

- Internal/external code security measures
- Separating the ML development environment from the rest of IT infrastructure
  - VMs/containers
  - Changing programming languages before deployment

Part III: Data minimisation and security

- Privacy attacks on ML models
  - model inversion attacks[1]
  - membership inference attacks
  - whitebox/blackbox attacks

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

- Mitigating the risks
  - assessing the training data if it contains identifiable personal data
  - avoiding overfitting in ML models
  - preventing blackbox attacks: monitoring API calls
  - preventing whitebox attacks: less control on the deployed model on the client-side

[1] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (*CCS '15*). ACM, 1322–1333.

Part III: Data minimisation and security

Data minimisation – Article 5(1)(c) of the GDPR

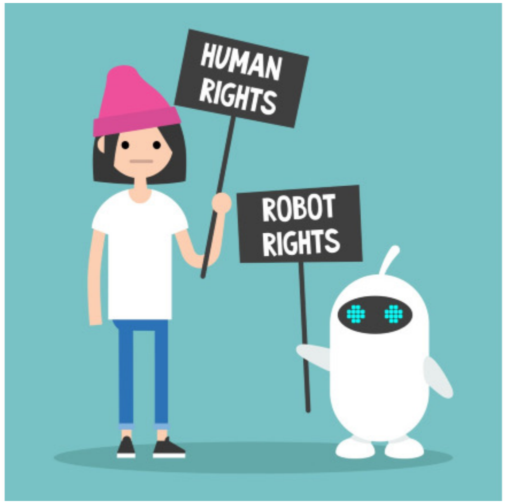
Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

Ensuring data minimisation:

- Training stage: Using feature selection techniques to select features which will be useful
- Training stage: Using privacy-enhancing methods (perturbation/adding noise and federated learning)
- Inference stage: less human-readable inputs, local inferences, privacy-preserving query approaches

Part IV: Rights in AI Systems

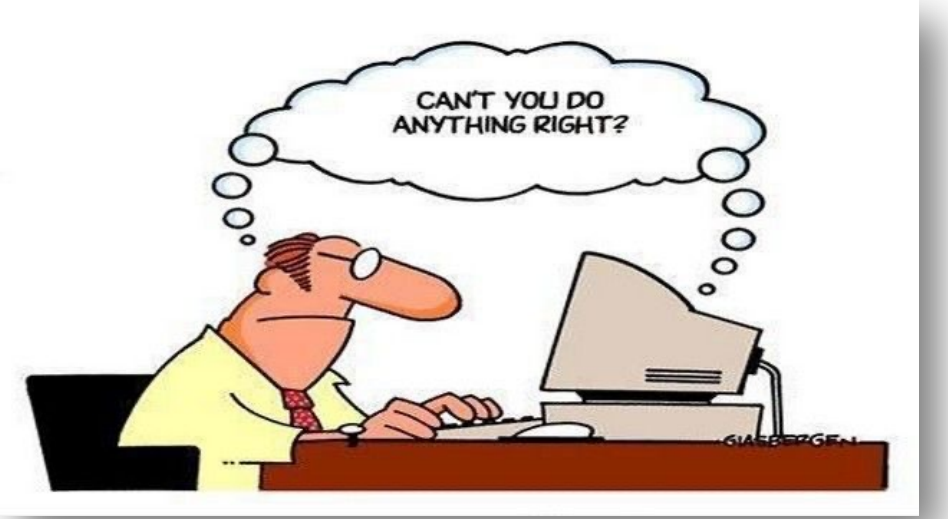
- individual rights requests for training data
  - right of access / rectification / erasure ('right to be forgotten') / data portability / being informed about the collection and use of their personal data
- individual rights requests for AI outputs
  - any model outputs that constitute personal data is subject to the rights of access, rectification, erasure
  - inferred personal data is out of scope of the right to portability



Part IV: Rights in AI Systems

- ensuring meaningful human input in non/partly automated decisions
  - requires training of staff

- ensuring meaningful human review of solely automated decisions
  - requires training of staff



Guidance Outline

- ICO is focusing on a risk-based approach to AI
  - Assessing the risks to the rights and freedoms of individuals that may arise
- Guidance prepared for:
  - an audience with a compliance focus (e.g., data
  -

More information in the guideline!

- Four main topics are covered:
  - Accountability and governance of AI
  - Fair, lawful, transparent processing
  - Data minimisation and security
  - Rights in AI systems
- Controls: Preventative, Detective, Corrective

Summary

- AI Auditing frameworks are becoming important.
- Internal auditing: SMACTR Framework
- External auditing: ICO Guidance
- Others: Non-profit organizations such as Algorithmic Justice League