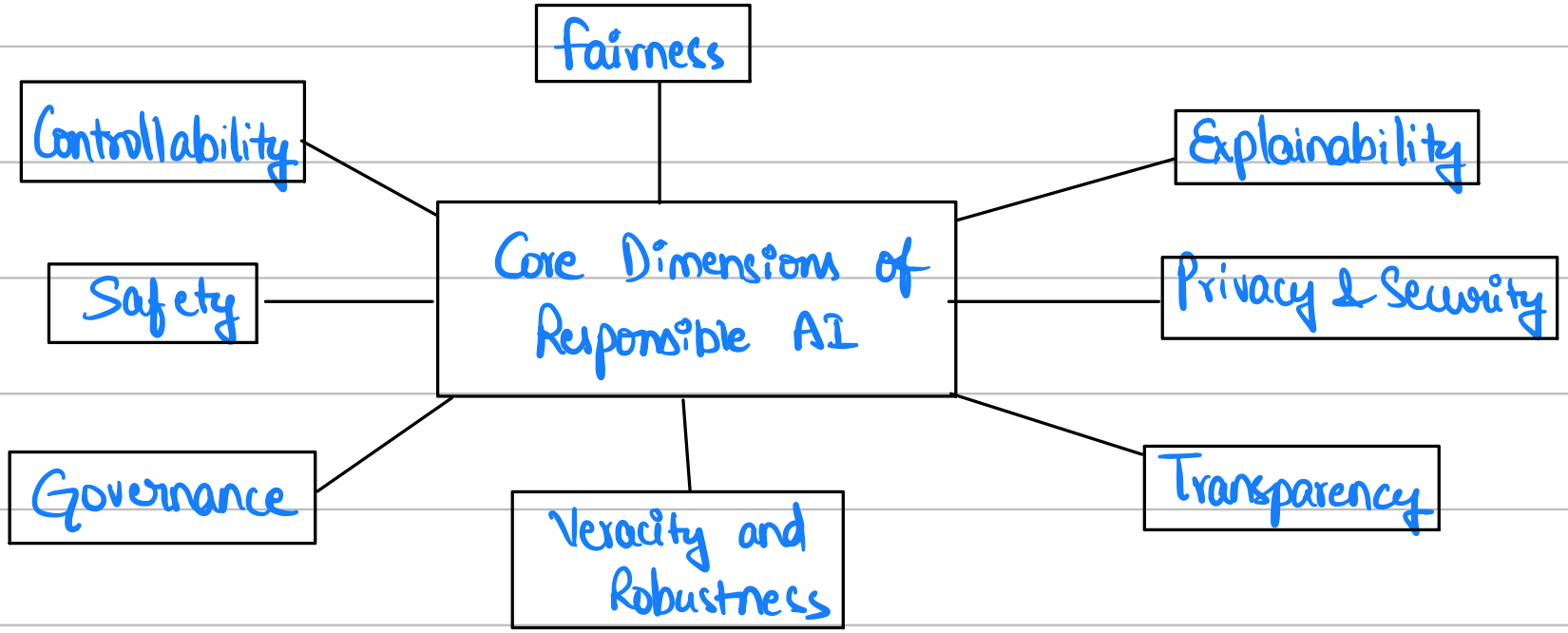
AI challenges and Responsibilities

- Responsible AI: To make sure that the AI systems are transparent & trustworthy while mitigating the risks and potential negative outcomes throughout the AI lifecycle.
  - Security: Ensure the confidentialy, integrity & availability are maintained on organizational data, information assets and infrastucture
  - Governance: Ensure to add value & manage risk in operation of business with clear policies & mechanisms to ensure AI system align with the legal and regulatory requirements. It's goal is to improve trust.
  - Compliance: To ensure adherence to regulations and guidelines regarding sensitive domains such as healthcare, finance and legal applications.



- The AWS Services for Responsible AI are:

- **Amazon Bedrock:** human or automatic model evaluation
- **Guardrails for Amazon Bedrock**
  - Filter content, redact PII, enhanced safety and privacy…
  - Block undesirable topics
  - Filter harmful content
- **SageMaker Clarify**
  - FM evaluation on accuracy, robustness, toxicity
  - Bias detection (ex: data skewed towards middle-aged people)
- **SageMaker Data Wrangler:** fix bias by balancing dataset
  - Ex: Augment the data (generate new instances of data for underrepresented groups)
- **SageMaker Model Monitor:** quality analysis in production
- **Amazon Augmented AI (A2I):** human review of ML predictions
- **Governance:** SageMaker Role Manager, Model Cards, Model Dashboard

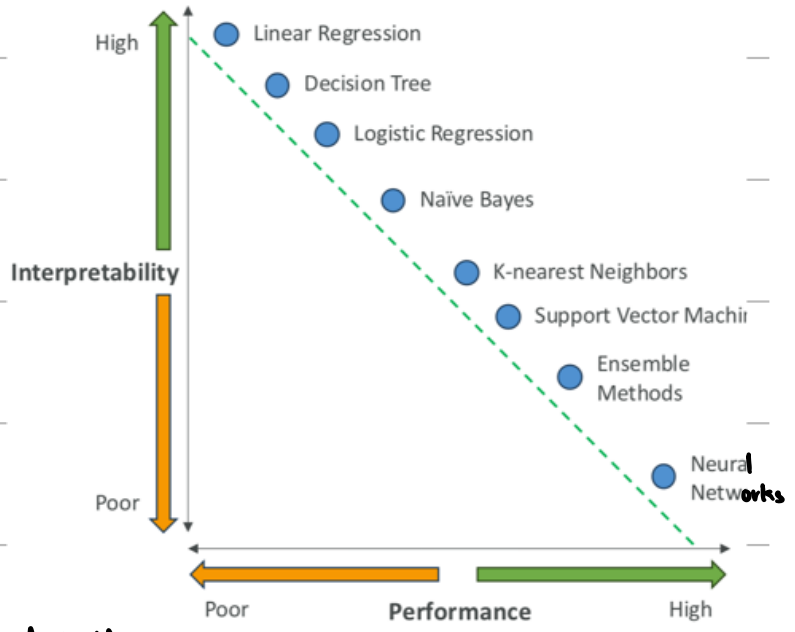
- AWS AI Service Cards: It is the documentation of the services. You can find out about the intended usecases, limitations and responsible AI design choices. They also contain deployment and performance optimization best practices.

- Interpretability Trade-offs:

↳ Interpretability is to answer why and how the model came to the decision.

High tranparency ⇒ High interpretability ⇒ poor performance

↳ Explainability is to understand the nature and behaviour of the model. See i/ps and o/ps and able to tell how the model came to conclusion without exact details.



- Partial Dependence Plots (PDP): Shows how a single feature can influence predicted

outcome, while holding other features constant. Useful when model is "blackbox" (NNs).

- Human-Centered Design (HCD) for Explainable AI:

↳ Approach to design AI systems with priorities for human needs.

↳ Design for amplified decision-making (stressful environment)

↳ Design for unbiased decision-making

↳ Design for human and AI learning.

Capabilities of GenAI	Challenges of GenAI
Adaptability	Regulatory Violations
Responsiveness	Social risks
Simplicity	Data Security & Privacy Concerns
Creativity & Exploration	Toxicity

Data Efficiency	Hallucinations
Personalization	Interpretability
Scalability	Non deterministic
	Plagiarism and Cheating

- Toxicity: Generating content that is offensive, disturbing or inappropriate. To mitigate this, we can curate training data or use guardrails.
- Hallucinations: Assertions or claims that sound true, but are incorrect. To overcome this, we have to verify the content generated by them.
- Plagiarism and Cheating: Worries that this technology is used to write college essays, or samples for job application or other forms of cheating.
- Prompt Misuses:

1. Poisoning is intentional introduction of malicious/biased data into training dataset.

2. Hijacking & prompt injection are used to influence the ops by embedding specific instructions within the prompts themselves.

3. Exposure is the risk of exposing sensitive or confidential information to a model during the training or inference.

4. Prompt Leaking is unintentional disclosure or leakage of the prompts and inputs used within the model.

5. Jailbreaking is to circumvent the constraints and safety measures implemented in a generative model to gain unauthorised access or functionality.

- AI Standard Compliance Challenges:

1. Complexity and Opacity

2. Dynamism and Adaptability

3. Emergent Capabilities

4. Unique Risks : Algorithmic Bias and Human Bias

5. Algorithm Accountability

- Model Cards: A standard format to document an ML model and it's key details.

- AWS Tools for Governance:

1. AWS Config
2. Amazon Inspector
3. AWS Audit Manager
4. AWS Artifact
5. AWS CloudTrail
6. AWS Trusted Advisor

- Governance Strategies

1. Policies
2. Review Cadence

3. Review Strategies
4. Transparency Standards
5. Team Training Requirements

- Data Governance Strategies:

1. Responsible AI
2. Governance Structure and Roles
3. Data Sharing and Collaboration

- Data Management Concepts:

1. Data Lifecycles
2. Data logging
3. Data Residency
4. Data Monitoring
5. Data Analysis
6. Data Retention

- Data Lineage:

1. Source Citation
2. Documenting Data Origins

3. Cataloging : organization & documentation of datasets.

- Security and Privacy for AI systems:

1. Threat Detection
2. Vulnerability Management
3. Infrastructure Protection
4. Prompt Injection
5. Data Encryption

- Monitoring AI systems:

1. Performance Metrics
2. Infrastructure Monitoring
3. Bias and fairness, Compliance and Responsible AI

AWS Responsibility - Security of the cloud

Customer Responsibility - Security in the cloud

Shared Controls - Patch & Configuration Mgmt. Awareness & Training

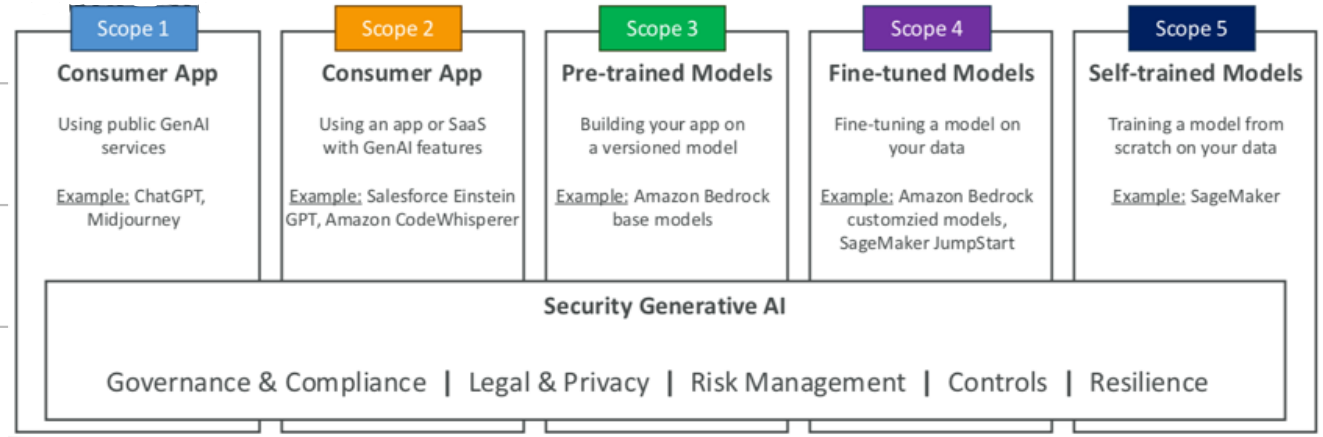
- Secure Data Engineering - Best Practices:

1. Accessing Data Quality
2. Privacy-Enhancing Technologies
3. Data Access Control
4. Data Integrity

- Generative AI Security Scoping Matrix:

↳ framework designed to identify and manage security risks associated with deploying GenAI applications.

↳ Classify your apps in 5 defined GenAI scopes from low to high ownership



- MLOps:

↳ Make sure the models are deployed, monitored, retrained systematically & repeatedly.

↳ The key principles include:

1. Version Control
2. Automation
3. Continuous Integration
4. Continuous Delivery
5. Continuous Retraining
6. Continuous Monitoring