Exploration & Practice on Generative AI Governance

2024.02

Juan Deng

Alibaba

云智能标准化部

CLOUD INTELLIGENCE STANDARDIZATION DEPARTMENT

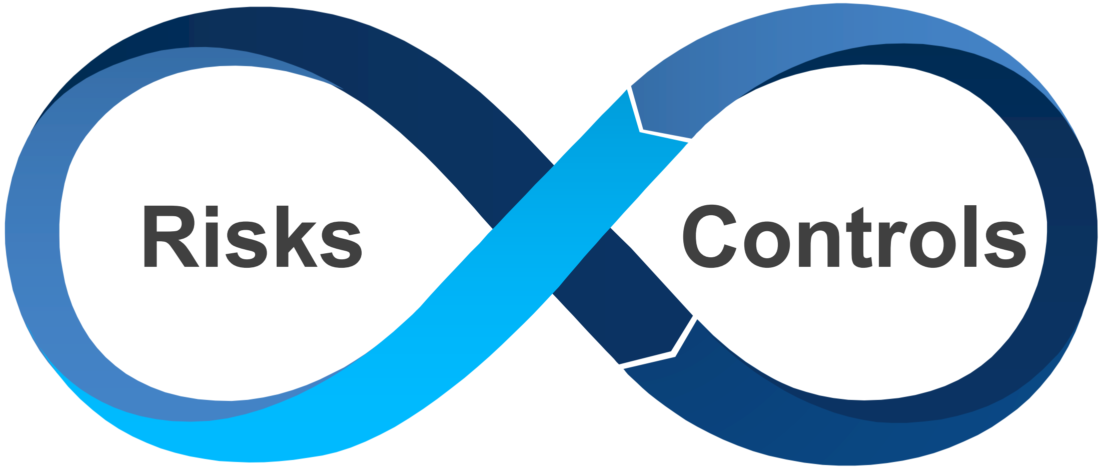
Contents

1 A Life-Cycle Approach

2 Risks of Generative AI

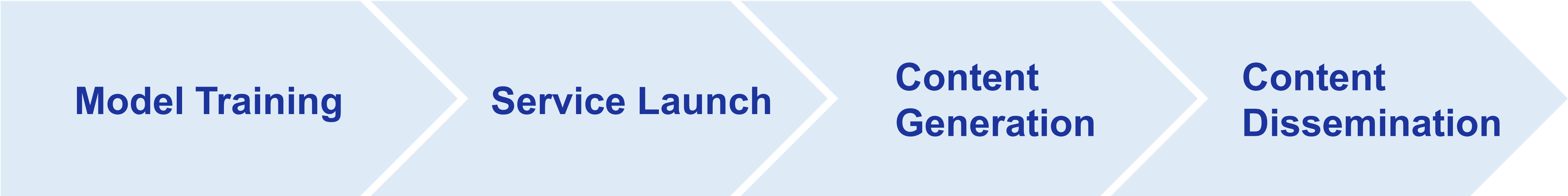
3 Controls

A Life-Cycle Approach



Risks

Controls



Model Training

Service Launch

Content Generation

Content Dissemination

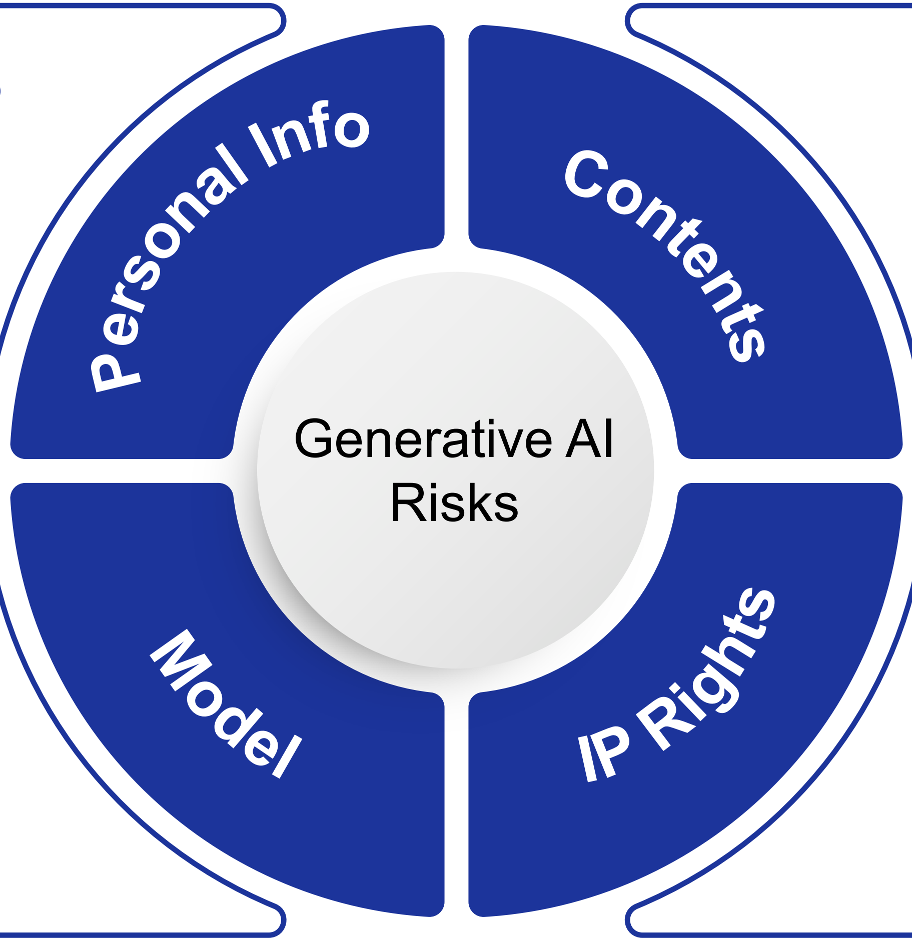
Risk Types of Generative AI

Concerns with personal info

- Personal info in training data, in input and output in real-time interaction
- unauthorized usage, cross-border issue…

Security issues of generated contents

- Illegal contents: contents that are pornographic, discriminatory, racist…
- False contents
- Contents against ethics, morals,…



Security issues of model

- Traditional software and info tech security issues: backdoor vulnerabilities, data theft, reverse engineering…
- Unfairness, adversarial attacks, inexplicability, data poisoning…

Intellectual property (IP) rights issues

- IP right infringement in training data
- Copyrightability of generated contents

Sources of Risks of Large Language Model

	Unsupervised pre-training	Supervised fine-tuning	Reinforcement learning	Content Generation
Personal Info	Training data may have unauthorized personal info	Training data may have unauthorized person info	Inappropriate reward model may encourage to output personal info	Users may use inducive dialogues
Content Security	Training data may contain illegal and harmful contents	Training data may contain illegal and harmful contents	Inappropriate reward model may amplify the issue	Users may use inducive dialogues
Model Security	Training data may contain biases	Training data may contain biases	Inappropriate reward model may amplify the issue	Users may use inducive dialogues
IP Rights	Training data may contain unauthorized contents	Training data may contain unauthorized contents	Inappropriate reward model may amplify the issue	User may submit unauthorized contents resulting in unauthorized generated contents

Sources of Risks of Large Vision Model

	Unsupervised pre-training	Generative Model Training	Model Alignment	Content Generation
Personal Info	Training data may have unauthorized personal info	Training data may have unauthorized person info	Personal info issues may become more prominent due to personalization	The generated content may contain biometric features
Content Security	Training data may contain illegal and harmful contents	Training data may contain illegal and harmful contents	Content security issues may become more prominent due to personalization	The generated content may illegal and harmful contents
Model Security	Training data may contain biases	The training data may contain biases	Customized models are more likely to be used maliciously	The generated content may be adversarial
IP Rights	Training data may contain unauthorized contents	Training data may contain unauthorized contents	Personalization leads to increased IP right issues, e.g. personalization of specific artistic styles	The generated content may contain artistic style info, style transfer, etc. leading to IP rights issues

Controls

	Model Training	Service Launch	Content Generation	Content Dissemination
Risk Source	• Data • Model	• Model • Tool-plugin	• User • Model	• Illegal and harmful info. • …
Controls	**Training Data** • Collection, filtering, labeling, review, etc. • Personal info protection **Model** • Benchmarks & evaluation • robustness/fairness considerations • …	**Model** • Benchmarks & evaluation **Tool-plugin** • Tool-plugin security **Stakeholders** • Model provider evaluation • User agreement • Transparency and information to users	**Content Review** • Content review mechanism • Content disposal • Person info collection, storage, use, deletion **User Management** • User account management	**Traceability** • Trace the publishers and disseminators **Operation** • Reporting, monitoring, etc…
General Requirement	Supply chain management and info storage for algorithmic services			

Thank you very much for your attentions！

云智能标准化部

CLOUD INTELLIGENCE STANDARDIZATION DEPARTMENT