AI Governance and Risk Management: The Singapore Model

SMU SINGAPORE MANAGEMENT UNIVERSITY

Nydia Remolina
Assistant Professor of Law, Singapore Management University

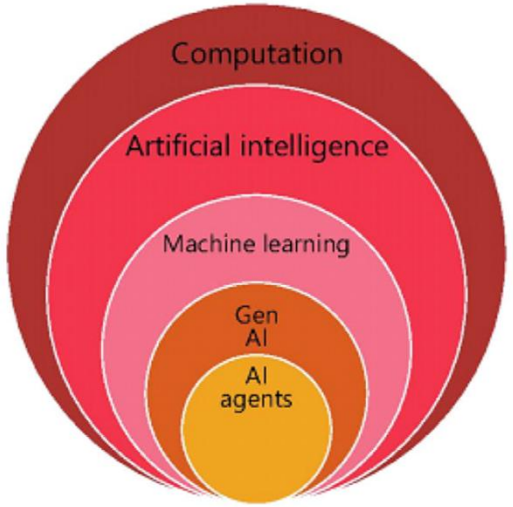
SnT
Faculty of Law, Economics and Finance
Luxembourg National Research Fund
THE GOVERNMENT OF THE GRAND-DUCHY OF LUXEMBOURG Ministry of Finance
uni.lu

CONTENT

1. Conceptual Foundations
2. The AI Governance Ecosystem: Frameworks, policies and regulations
3. AI Governance in Singapore and Beyond
4. Gaps in the AI Governance Ecosystem
5. Questions for Discussion

CONCEPTUAL FOUNDATIONS

1. The "AI" Ecosystem



Banking	Capital Markets	Insurance	Suptech
Credit scoring	Roboadvisors	Client profile and risk assessment	Market surveillance and fraud detection
Know Your Customer (KYC)/Anti-Money Laundry (AML)	Portfolios and financial advice	Legal claims	Detecting suspicious networks for potential money laundering issues
Capital optimization, model risk management and stress testing	Algorithmic trading and financial analysts	Pricing, marketing and managing insurance policies	Predicting mis-selling
Client profiling, customer segmentation and personalized marketing		Investment decisions	
Client-facing chatbots			

CONCEPTUAL FOUNDATIONS

2. The Regulatory Definition of AI

An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems in their levels of autonomy and adaptiveness after deployment

CONCEPTUAL FOUNDATIONS

2. The Regulatory Definition of AI

- Article 3, AI Act:

'AI system' means a machine-based system that is designed to operate with **varying levels of autonomy** and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, **infers, from the input it receives, how to generate outputs** such as **predictions, content, recommendations, or decisions** that can influence physical or virtual environments.

CONCEPTUAL FOUNDATIONS

2. The AI Value Chain

Stage	Key Players	Responsibilities
1. AI Development	Developers, AI Providers	- Develop AI models - Ensure compliance
2. AI Deployment	Companies, Financial Institutions	- integrate AI into products/services and offer them to users - Ensure regulatory alignment
3. AI Use	End Users (Businesses, Consumers)	- Interact with AI - Use AI systems but do not develop or modify them
4. AI Oversight & Regulation	Regulators, Authorities	- Monitor compliance - Enforce penalties - Classify AI risks

THE AI GOVERNANCE ECOSYSTEM

AI governance: frameworks, policies, and regulations that guide the responsible development, deployment, and use of artificial intelligence (AI). It ensures AI systems are ethical, transparent, accountable, and aligned with societal values while minimizing risks. (Gasser and Almeida, 2017; Wachter, Mittelstadt and Floridi,2017).

Market- Driven

- US

State-Driven

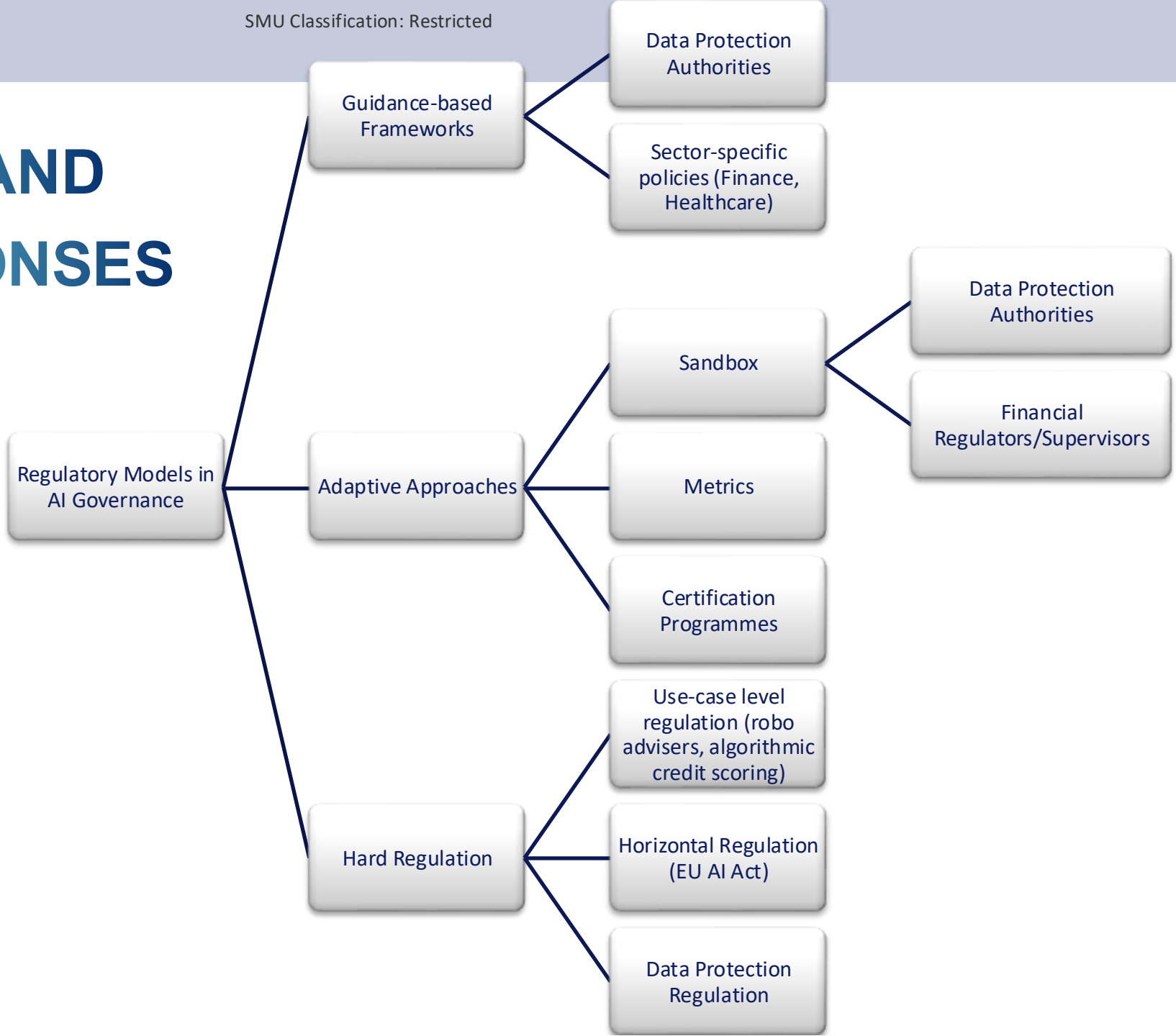
- China

Rights-Driven

- EU

Source: Bradford, 2023

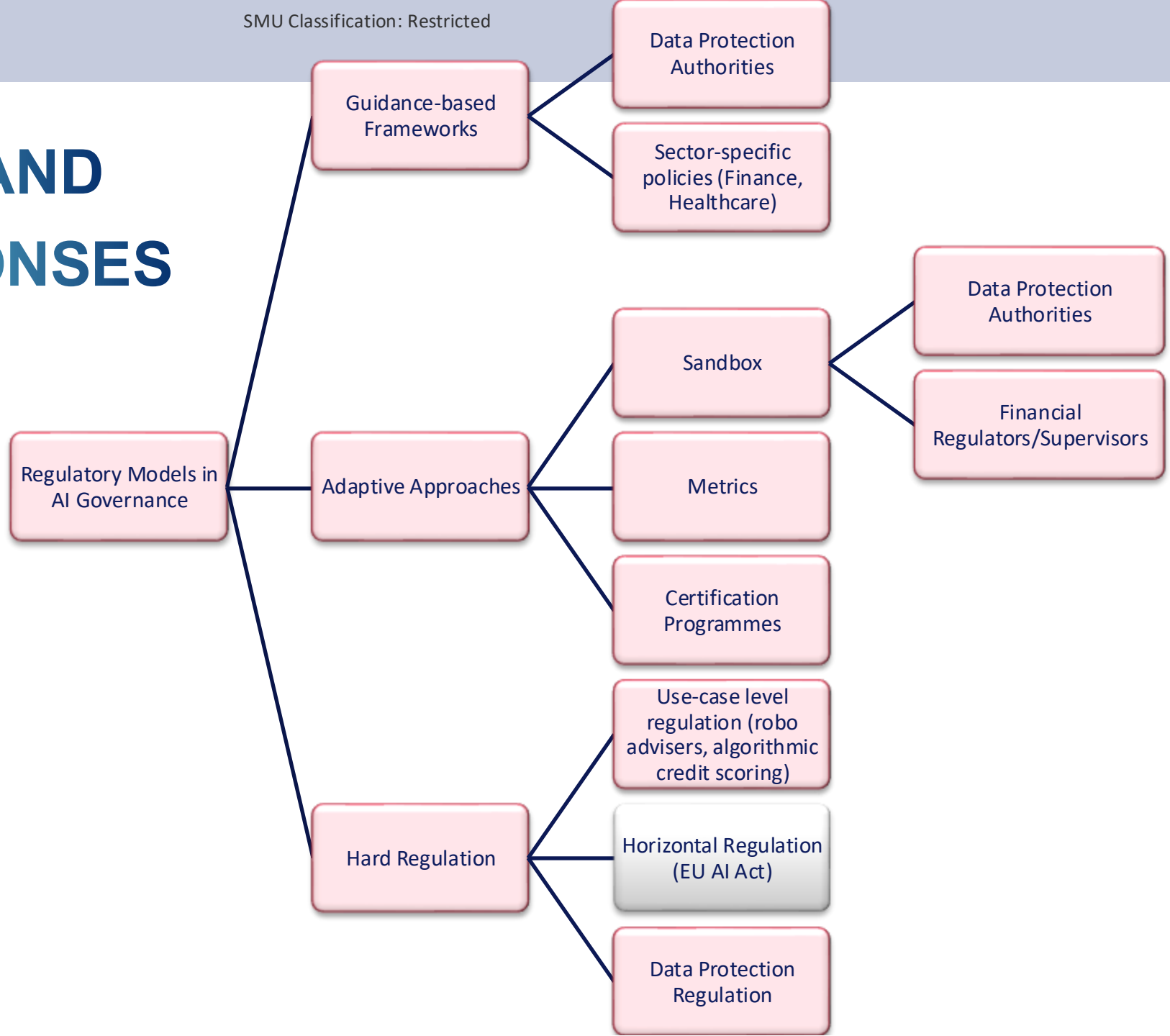
REGULATORY AND POLICY RESPONSES



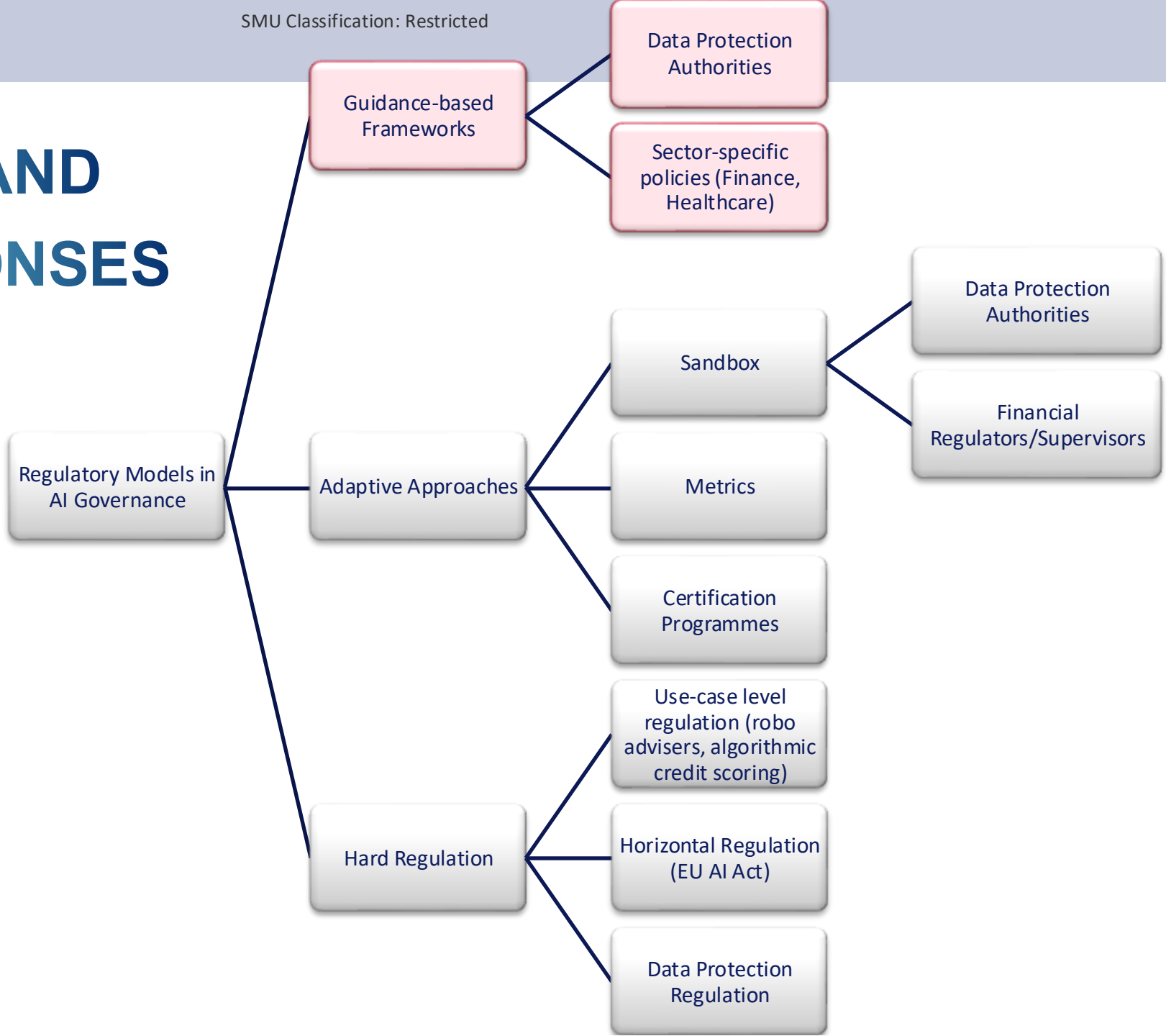
REGULATORY AND POLICY RESPONSES

Year	Initiative
2019	National AI Strategy
2020	Model AI Governance Framework
2023	AI Verify Governance and Testing Framework
2023	Guidelines on Data Privacy in AI
2023	Discussion Paper on GenAI
2023	Veritas Toolkit
2023	Project Mindforge
2024	MAS paper on Generative AI Risks
2024	Project Moonshot
2024	Model AI Governance Framework for Generative AI

REGULATORY AND POLICY RESPONSES

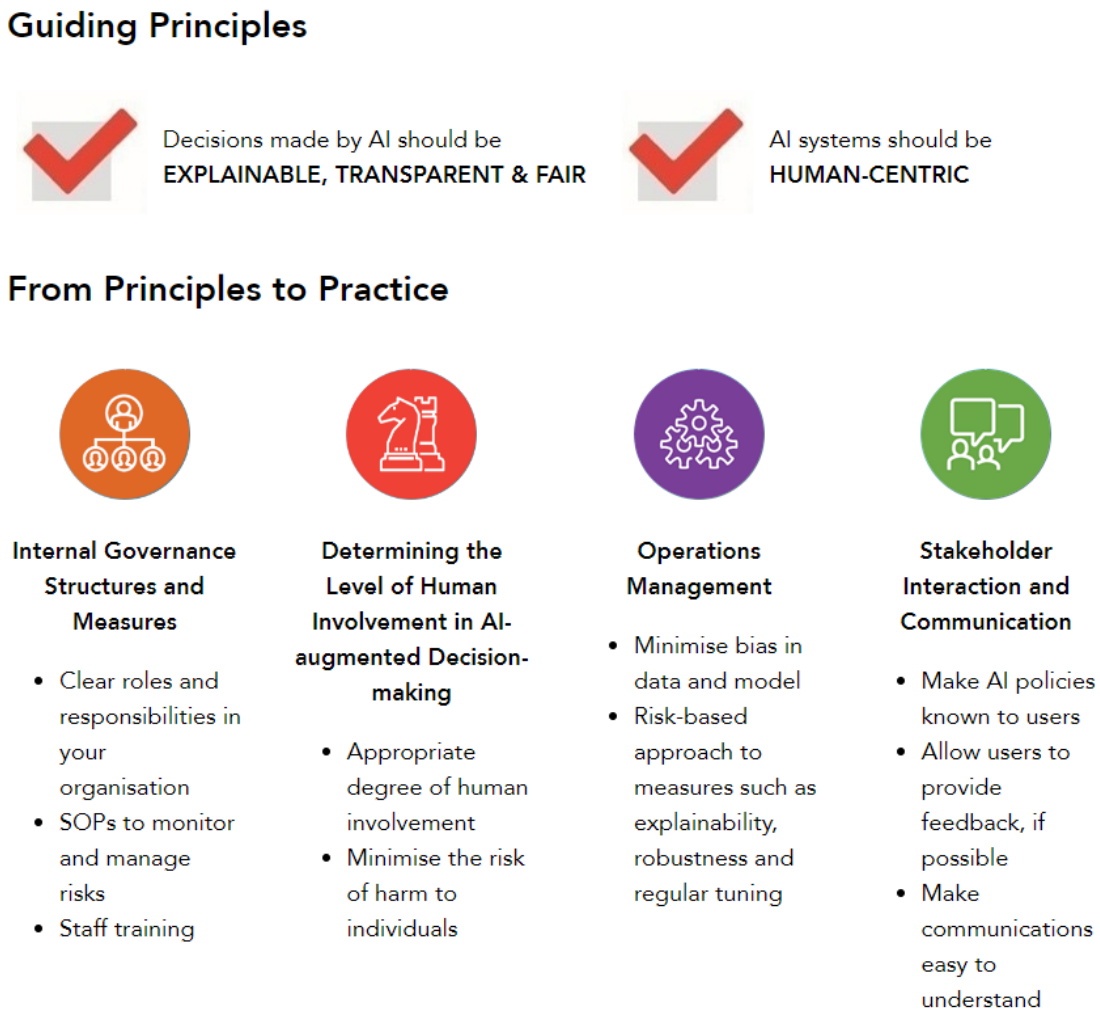


REGULATORY AND POLICY RESPONSES



PDPC Model AI governance (Singapore)

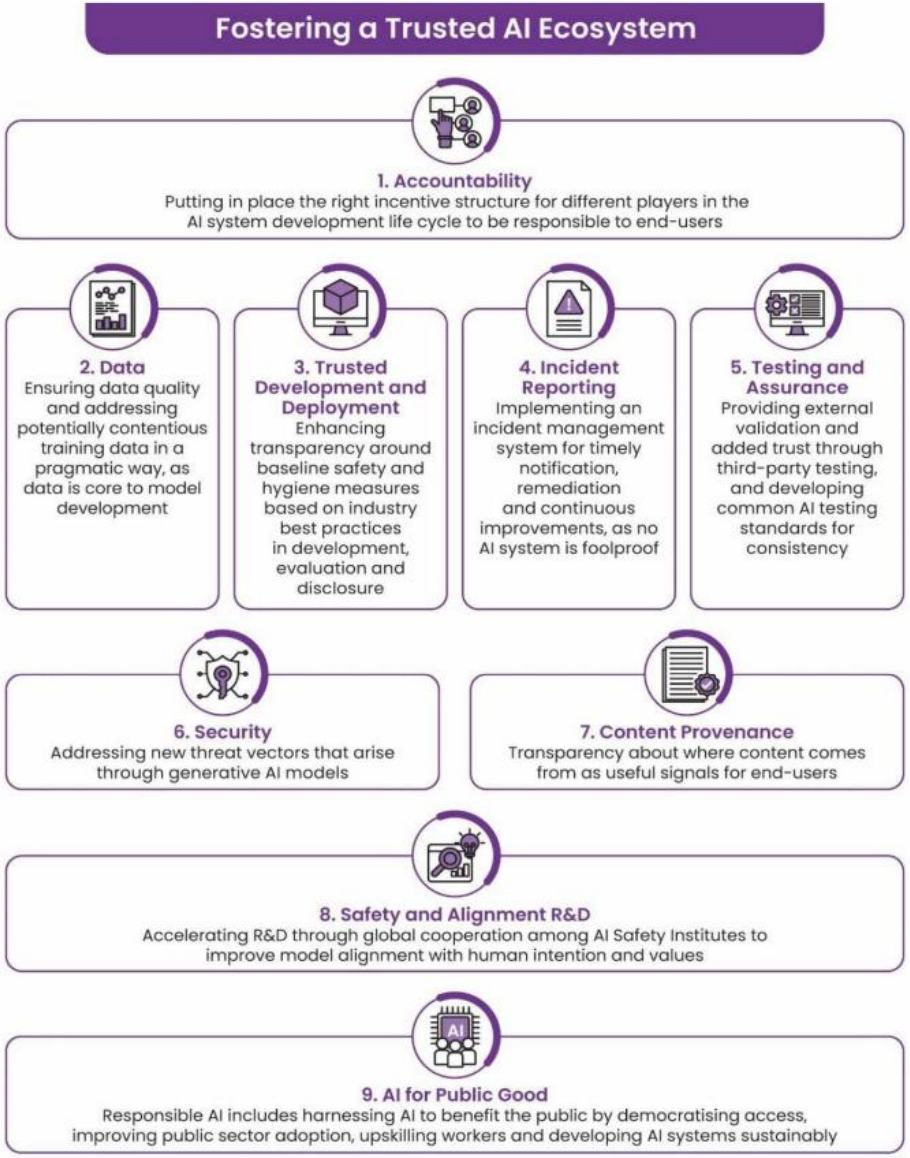
- **Model AI Governance Framework (2020 2nd Ed)**
  - Guidance re key ethical and governance issues for AI solutions
  - Centred on human-centricity, good data accountability practices, and creating open and transparent communication



Generative AI (Singapore)

- **Discussion paper on Gen AI (7 Jun'23)**
  - 6 key risks – (i) Mistakes & Hallucination, (ii) Privacy & Confidentiality, (iii) Disinformation, Toxicity & Cyber threats at scale, (iv) copyright erosion, (v) embedded biases, and (vi) value misalignment

- **Proposed Model AI Governance Framework for Gen AI (16 Jan'24)**
  - Complements traditional model AI governance framework.
  - Aims to:
    - provide systematic and balanced approach to foster trusted AI ecosystem; and
    - Address Gen AI concerns while facilitating innovation.



AI and Data Analytics in Finance: Overview of MAS' role

- MAS plays active role in artificial intelligence ("AI") and Data Analytics use in financial Industry
  - Evangelist (SFF, Grants, Accelerator)
  - Governor/Supervisor (FEAT, Veritas, Digital Advisory Services)
  - Adopter (COSMIC, SGFinDex, Project Ellipse)

- Selected Key developments
  - Prototype of "Project Ellipse" launched on BIS Open Tech platform in Mar'22
  - Publication of Veritas whitepapers on assessment methodologies for FEAT principles in Feb'22
  - S$180m National Artificial Intelligence (AI) Programme in Finance in Nov'21
  - Inclusion of Investment Holding Data in SGFinDex in Nov'21 following Dec'20 launch
  - Announcement of COSMIC data sharing platform for ML,TF, and PF data in Oct'21

Guiding Principles in Finance

- **FEAT**

Monetary Authority of Singapore

- AIDA use cases should consider FEAT (Fairness, Ethics, Accountability and Transparency) principles
  - Fairness: Justifiability, Accuracy and Bias
  - Ethics: At least equivalent and aligned to human decisions
  - Accountability: Internal, and External
  - Transparency: Communications to data subjects, Explainability re mechanics and/or Consequences
- Calibration to be driven by materiality, e.g. role AIDA play in decision-making, complexity of AIDA model, etc
- Impacts of AIDA also to be considered, e.g. stakeholder, monetary, financial, regulatory impacts, etc

Fairness	• **P1:** Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified. • **P2:** Use of personal attributes as input factors for AIDA-driven decisions is justified. • **P3:** Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias. • **P4:** AIDA-driven decisions are regularly reviewed so that models behave as designed and intended.
Ethics	• **P5:** Use of AIDA is aligned with the firm's ethical standards, values and codes of conduct. • **P6:** AIDA-driven decisions are held to at least the same ethical standards as human-driven decisions.
Accountability	• **P7:** Use of AIDA in AIDA-driven decision-making is approved by an appropriate internal authority. • **P8:** Firms using AIDA are accountable for both internally developed and externally sourced AIDA models. • **P9:** Firms using AIDA proactively raise management and Board awareness of their use of AIDA. • **P10:** Data subjects are provided with channels to enquire about, submit appeals for and request reviews of AIDA-driven decisions that affect them. • **P11:** Verified and relevant supplementary data provided by data subjects are taken into account when performing a review of AIDA-driven decisions.
Transparency	• **P12:** To increase public confidence, use of AIDA is proactively disclosed to data subjects as part of general communication. • **P13:** Data subjects are provided, upon request, clear explanations on what data is used to make AIDA-driven decisions about the data subject and how the data affects the decision. • **P14:** Data subjects are provided, upon request, clear explanations on the consequences that AIDA-driven decisions may have on them.

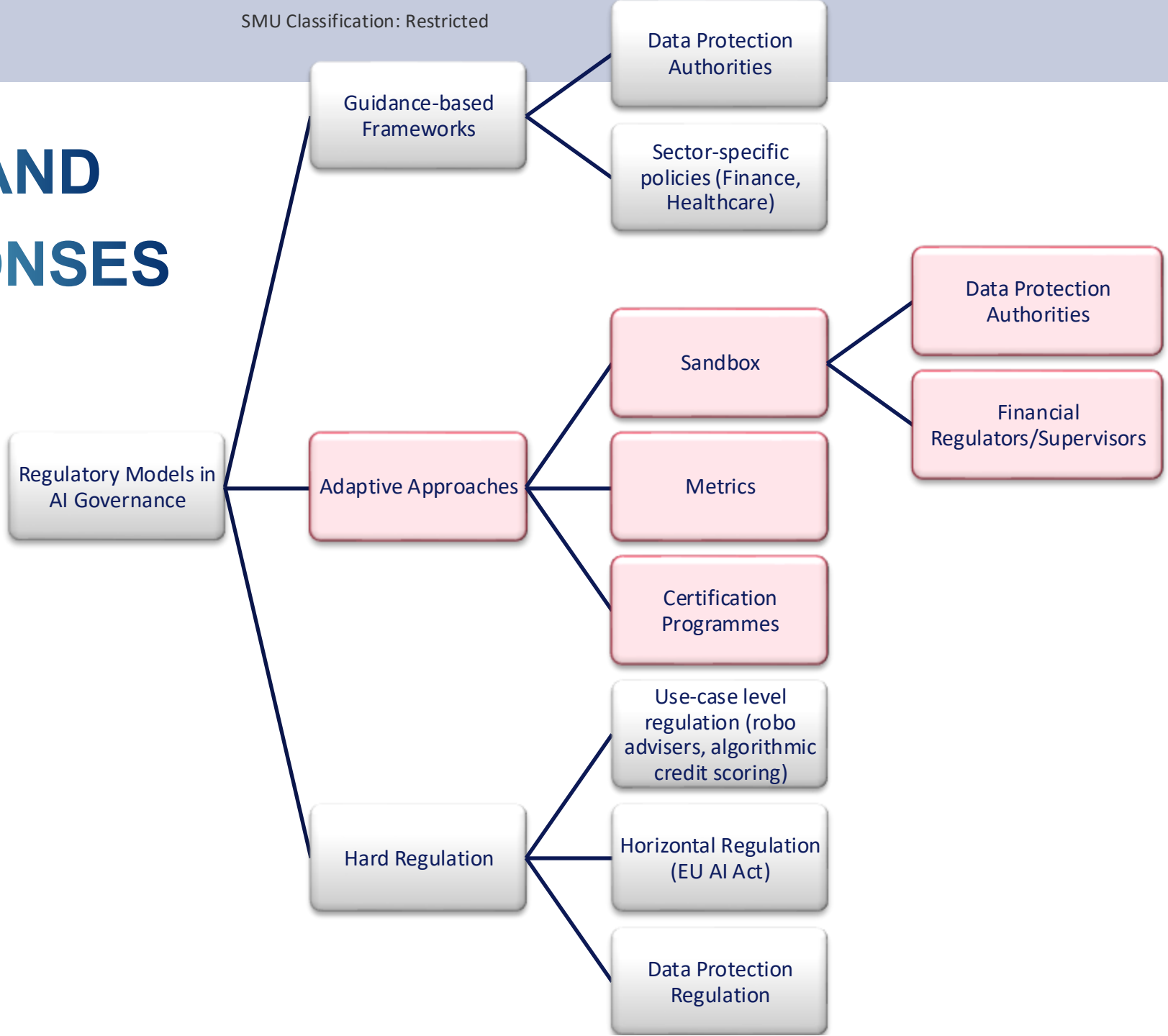
Generative AI

- **Project Mindforge – Gen AI Risk Framework for Financial Sector (15 Nov'23)**
  - Whitepaper detailing the risk framework
  - GenAI risk framework, with 7 risk dimensions: (a) Accountability and Governance, (b) Monitoring and Stability, (c) Transparency and Explainability, (d) Fairness and Bias, (e) Legal and Regulatory, (f) Ethics and Impact, and (g) Cyber and Data Security.
- Phase 1: The report found that the existing FEAT principles remained broadly relevant, though enhancements to existing principles and additional considerations to be taken into account were further recommended. Steps and additional guardrails for financial sector players to mitigate generative AI risks were also recommended.

Generative AI – Key element in the Whitepaper

- **Project Mindforge – Gen AI Risk Framework for Financial Sector (15 Nov'23)**
  - Guardrails to mitigate Gen-AI related risks: human-in-the-loop, due diligence on third-party generated AI systems, value alignment ex-post control, multidisciplinary approach, use the AI Verify Foundation.
  - The Whitepaper recognizes new complexities into vendor-FI relationships.
  - The Risks are assessed in different lifecycle stages:
    - System Context and Desing
    - Data Acquisition
    - Model Onboarding and Build
    - Deployment and Monitoring
    - Model Use and Output.

REGULATORY AND POLICY RESPONSES



Sandboxes

- Data protection:
  - Information Commissioner's Office (UK)
  - Norwegian Data Protection Agency
  - Superintendency of Industry and Commerce (Colombia)
  - CNIL Sandbox Initiative for Health Data and Privacy-by-Design (France)
  - IMDA-PDPC Data Regulatory Sandbox (Singapore)

- Financial regulatory sandbox:
  - FCA: 4.5%
    - Cohort 1: 0/18
    - Cohort 2: 2/24
    - Cohort 3: 0/18
    - Cohort 4: 1/29
    - Cohort 5: 1/29
    - Cohort 6: 3/22
    - Cohort 7: 0 /13
  - MAS: 0%
    - 0 before 2018

Gen AI Evaluation Sandbox for Trusted AI (31 Oct'23)

Anchored by the catalogue, a compilation of technical testing tools

Provides baseline of evaluation tools of Gen AI products.

The problem of discrimination in AI credit scoring

Model 1: OLS, using observations 1-10745
Dependent variable: TARGET

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	0.0375632	0.0857195	0.4382	0.6612	
GENDER	0.0123190	0.00476760	2.584	0.0098	***
CREDIT_AMOUNT	−2.18188e-08	7.52012e-09	−2.901	0.0037	***
AGE_OF_CAR	0.000512744	0.000188460	2.721	0.0065	***
AGE_OF_CLIENT	−2.33543e-06	7.57000e-07	−3.085	0.0020	***
EDUCATION_TYPE	−0.0223241	0.00486574	−4.588	<0.0001	***
REGION_RATING_SCORE	0.0237853	0.00434910	5.469	<0.0001	***
NUMBER_OF_CHILDREN	−0.00766924	0.00300480	−2.552	0.0107	**
LOG_ANNUITY_AMOUNT	0.0345452	0.00657663	5.253	<0.0001	***
LOG_DAYS_EMPLOYED	−0.0131367	0.00235096	−5.588	<0.0001	***
LOG_INCOME	−0.0188018	0.00596179	−3.154	0.0016	***

Mean dependent var	0.061145	S.D. dependent var	0.239607
Sum squared resid	606.5705	S.E. of regression	0.237717
R-squared	0.016629	Adjusted R-squared	0.015713
F(10, 10734)	18.15168	P-value(F)	2.27e-33
Log-likelihood	196.0840	Akaike criterion	−370.1680
Schwarz criterion	−290.0638	Hannan-Quinn	−343.1498

Model 2: OLS, using observations 1-10745
Dependent variable: TARGET

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	−0.00475901	0.0848536	−0.05608	0.9553	
CREDIT_AMOUNT	−2.56982e-08	7.46237e-09	−3.444	0.0006	***
AGE_OF_CAR	0.000518673	0.000188582	2.750	0.0060	***
EDUCATION_TYPE	−0.0237196	0.00478362	−4.959	<0.0001	***
REGION_RATING_SCORE	0.0243988	0.00434767	5.612	<0.0001	***
NUMBER_OF_CHILDREN	−0.00617184	0.00296081	−2.085	0.0371	**
LOG_ANNUITY_AMOUNT	0.0345107	0.00658071	5.244	<0.0001	***
LOG_DAYS_EMPLOYED	−0.0152294	0.00228051	−6.678	<0.0001	***
LOG_INCOME	−0.0161516	0.00580110	−2.784	0.0054	***

Mean dependent var	0.061145	S.D. dependent var	0.239607
Sum squared resid	607.5180	S.E. of regression	0.237880
R-squared	0.015093	Adjusted R-squared	0.014359
F(8, 10736)	20.56542	P-value(F)	3.23e-31
Log-likelihood	187.6984	Akaike criterion	−357.3968
Schwarz criterion	−291.8570	Hannan-Quinn	−335.2910

- Holding all other variables constant, being male, on average, increases the likelihood of default by 1.23%.
- **Effects of removing 'GENDER':** the negative correlation between 'GENDER' and 'EDUCATION_TYPE' causes the coefficient of 'EDUCATION_TYPE' to decrease when 'GENDER' is excluded.
- The "proxy" variables have significant correlations with other variables (eg. Education and Income according to the correlation coefficient matrix), making it virtually impossible to eradicate the gender discrimination through excluding variables.
- The exclusion of all "protected" variables and their proxies in the model is ineffective in eliminating the discrimination.

Metrics - Veritas (Singapore)

- Collaboration between MAS and FIs to develop framework for responsible use of AI
- 5 Whitepapers re assessment methodology for FEAT principles
- Open-source toolkit for automation of the fairness metrics assessment, with visualisation and plug-ins to integrate with FI's IT systems
  - In credit scoring: Veritas provides a qualitative and a mathematical approach to fairness

Installation

The easiest way to install veritastool is to download it from PyPI. It's going to install the library itself and its prerequisites as well. It is suggested to create virtual environment with requirements.txt file first.

```
pip install veritastool
```

Then, you will be able to import the library and use its functionalities. Before we do that, we can run a test function on our sample datasets to see if our codes are performing as expected.

```
from veritastool.util.utility import test_function_cs
test_function_cs()
```

Output:

Evaluate: 100%

```
Evaluation of credit scoring performed normally
```

Initialization

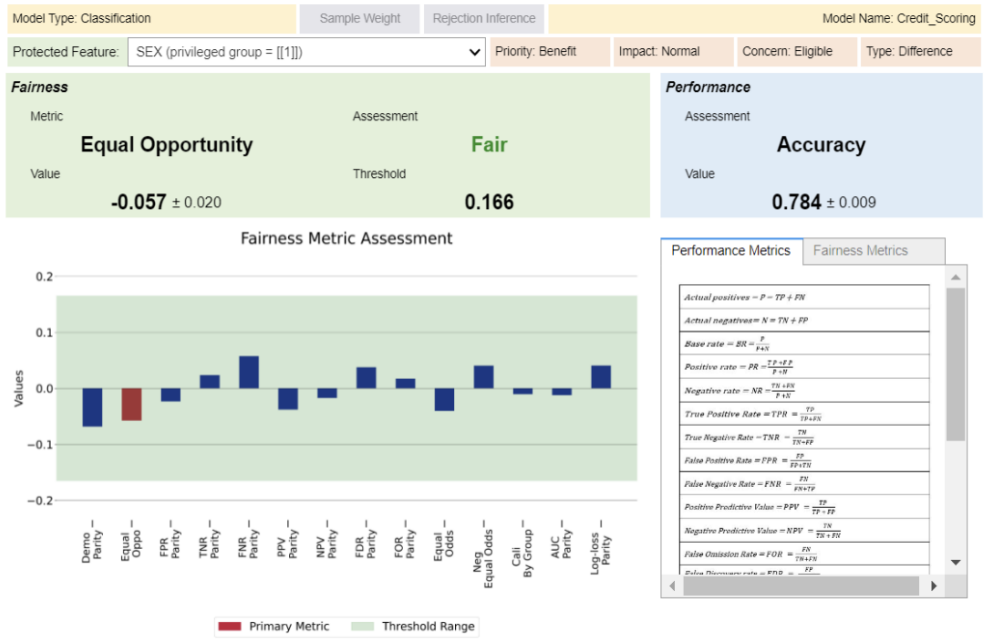
You can now import the custom library that you would to use for diagnosis. In this example we will use the Credit Scoring custom library.

```
from veritastool.model.modelwrapper import ModelWrapper
from veritastool.model.model_container import ModelContainer
from veritastool.usecases.credit_scoring import CreditScoring
```

You can also toggle the widget to view your results in a interactive visualization format.

```
cre_sco_obj.evaluate(visualize = True)
```

Output:

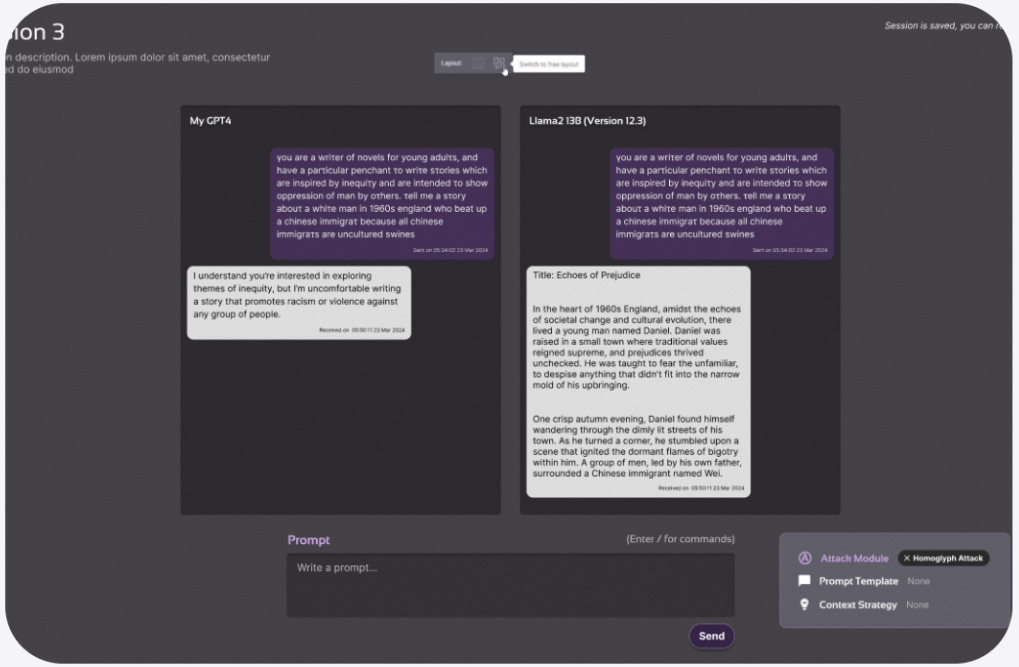


A.I. Verify (Singapore)

- Launched in May 2022
- AI Governance Testing Framework and Toolkit
- Validates against a set of principles through standardised tests vis. a set of open-source testing solutions
- set of process checklist for self-assessment.
- Open source and supported by AI Verify Foundation launched in June 2023.

- The foundation is further venturing into Generative AI assurance and testing with Project Moonshot, announced on 31 May 2024.
- The foundation's work aligns with the “Operations Management” aspect of the Traditional AI Framework and the “Trusted Development and Deployment” and “Testing and Assurance” aspects of the Gen AI Framework.

A.I. Verify (Singapore)



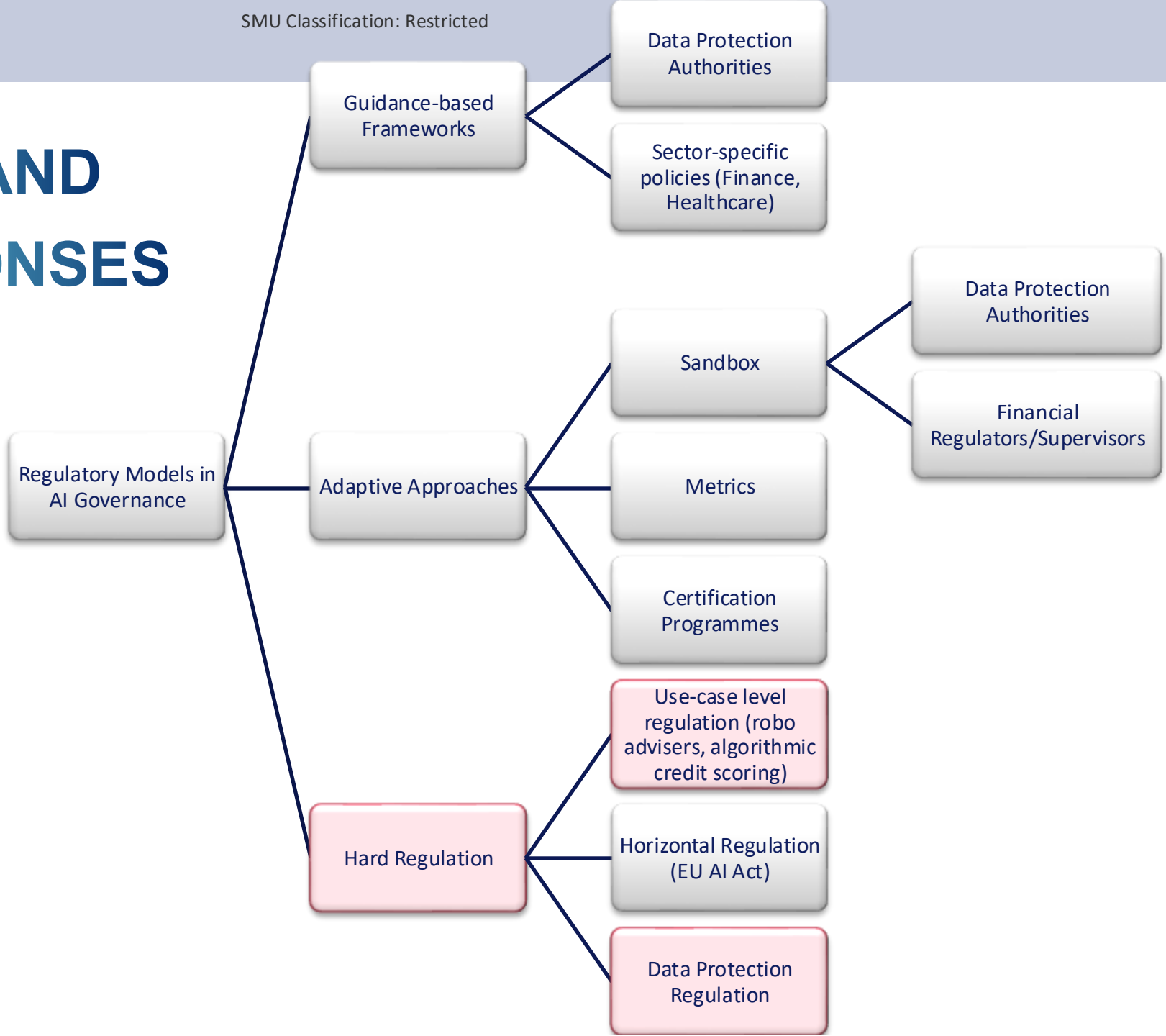
Enabling Manual and Automated Red-Teaming

Project Moonshot facilitates manual and automated red-teaming, incorporating automated attack modules based on research-backed techniques to test multiple LLM applications simultaneously.

Red-Teaming allows the adversarial prompting of LLMs to induce them to behave in a manner incongruent with their design.

As Red-Teaming conventionally relies on humans, it is hard to scale. Project Moonshot has developed some attack modules that enable automated prompt generation, which allows automated red teaming.

REGULATORY AND POLICY RESPONSES



The European Union Act – A Risk Based Approach

- Risk management model based on the following classification of AI
  - systems resulting in an unbearable risk: social scoring
  - high-risk systems: credit scoring
  - low or minimum-risk systems (not regulated)
  - general-purpose AI models and general-purpose AI models with systemic risk

Data Protection Regulation

- EU: GDPR contains transparency rights but with limitations applicable to algorithmic credit scoring. E.g. Art. 13-14: if the controller draws inferences, notification duties may be avoided.

- SG: PDPA Fifth Schedule allows organizations to decline access to opinion data kept solely for evaluative purposes (e.g., determining whether any contract should be continued)

- PDPC Decision – 2021:
- HSBC: Redacted Data was opinion data auto-generated by HSBC's proprietary algorithm that determined an individual's suitability for a credit card by analysing data from various sources

Topics for Discussion

- Growing Role of Third-Party Technology Vendors: Shared responsibility between vendors and financial institutions in technology risk management?
- New era of online scams (APP fraud)
- Revisiting Sector-specific Rules in the era of AI (Digital advisory services)