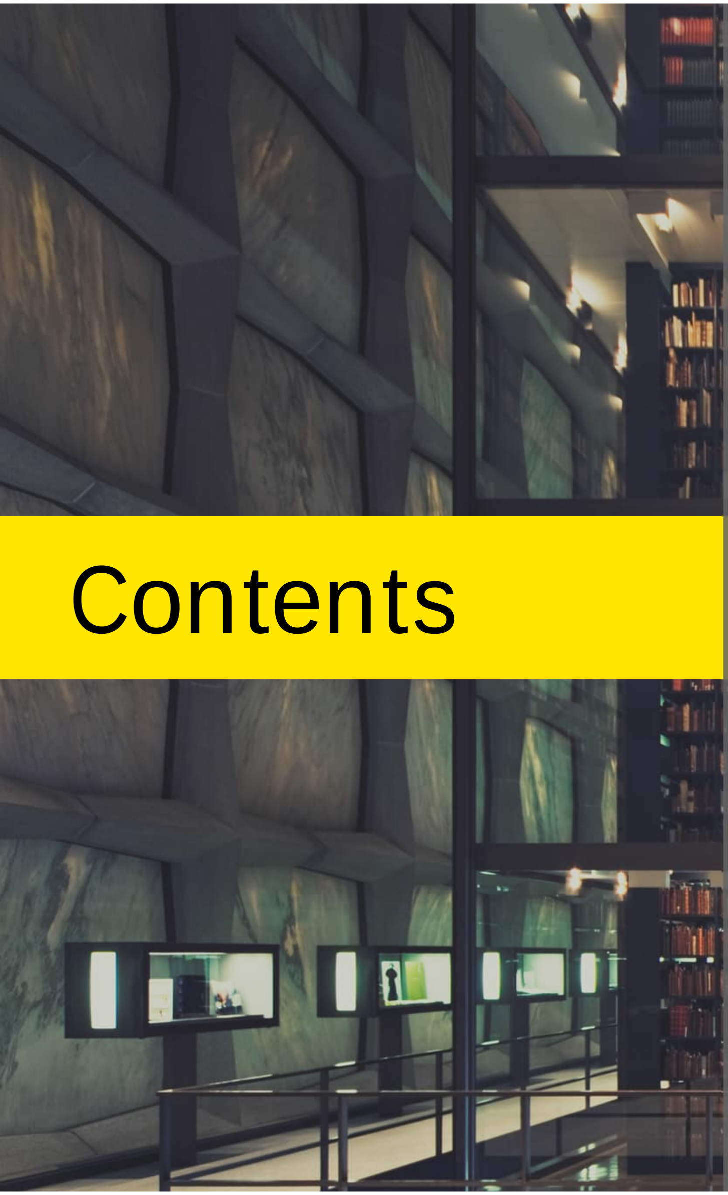
AI Security & Privacy

Way towards a responsible & trusted AI

Hitender Saxena

EY

Building a better working world



Contents

1 Introduction to AI

*What is AI
*Forecast
*AI & Cloud Adaption
*Common AI Workloads
*Principals of AI

2 AI & Security

*Three dimensions of AI security:
*Security Risks Of AI
*Security Applications of AI
*AI Security Management
*AI - Attack Methods & Impact

3 Hidden Dangers of AI

*Algorithm Model
*Data Security & Privacy Protection
*Infrastructure
*Application
*Abuse of AI

4 How EY Can Help?

*Key risks associated with AI
*EY's Trusted AI Framework
*EY's Trusted AI Lifecycle
*GRC

5 Conclusion

6 References

Introduction to Artificial Intelligence

What is Artificial Intelligence?

Simulation of human intelligence done by machines programmed by us. AI is the application of **Statistic, Machine learning and Robotics** to deliver **Prediction, Automation or Augmentation** tools

Prediction

Application of ML & Statistics to predict answers to business problems E.g. Fraud detection

Automation

Application of Robotics & ML to automate business tasks E.g. Automated invoice recognition & processing

Augmentation

Application of tools incorporating prediction and automation to augment a human's capacity to do their job. E.g. Augmented Sales team

Simply put, AI is the creation of software that imitates human behaviours and capabilities. Key elements include:

- Making decisions based on data and past experience
- Detecting anomalies
- Interpreting visual input
- Understanding written and spoken language
- Engaging in dialogs and conversations

AI and Cloud Adoption

- AI, IoT and the cloud go together in today’s technological ventures.
- Digital assistants like Apple Siri, Google Home and Amazon Alexa have penetrated every aspect of our lives and were created using artificial intelligence methods and cloud resources.
- Tasks such as ordering online, using a household fixture/appliance, making an appointment, listening to music, asking a question, and even communicating with someone over text or calling them directly can now be done using digital assistants.
- Ability to scale operations in an effective and efficient manner. Computing resources can be replicated with a click of a button to scale up or down as needed.
- CSP’s provide pre-trained and ready to use machine learning, deep learning and other artificial intelligence models, algorithms and services for businesses to use in their data analytics process.
- Access powerful models that have been trained on millions and even billions of rows of data at a fraction of the cost.

Common Artificial Intelligence Workloads

Machine learning

This is often the foundation for an AI system, and is the way we “teach” a computer model to make prediction and draw conclusions from data.

Anomaly detection

The capability to automatically detect errors or unusual activity in a system.

Computer vision

The capability of software to interpret the world visually through cameras, video, and images.

Natural language processing

The capability for a computer to interpret written or spoken language, and respond in kind.

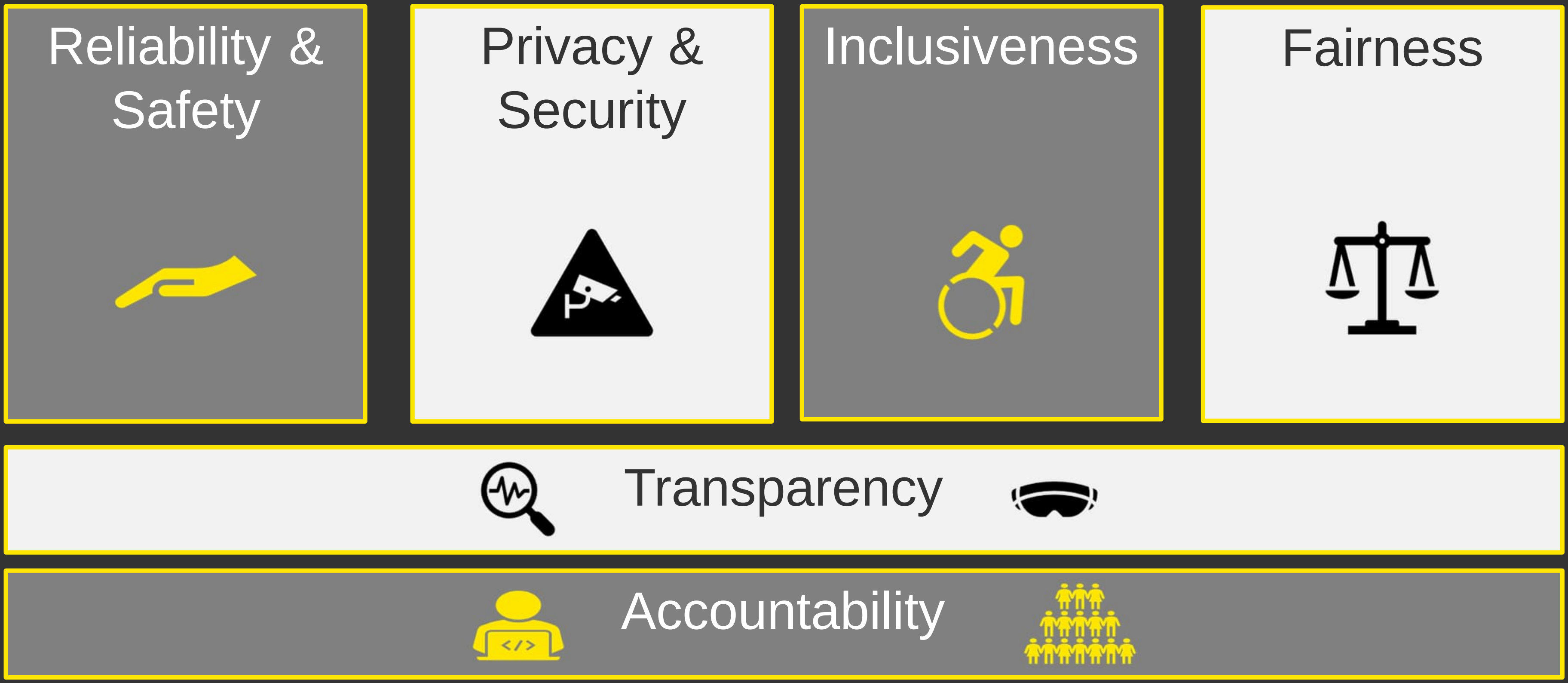
Conversational AI

The capability of a software agent (usually referred to as a bot) to participate in a conversation.

AI-related workloads

Principles of Responsible AI

Human + AI



Reliability & Safety

Privacy & Security

Inclusiveness

Fairness

Transparency

Accountability

AI and Security

Three dimensions of AI security:

- Reduce AI immaturity and the security risks malicious applications pose to cyberspace and national society
- Promote the deep application of AI in the fields of cybersecurity and public safety
- Establish an AI security management system to ensure the safe and steady development of AI.

Risk

Negative impact of AI technology and industry on cyberspace security and national societal security.

Application

Specific application directions of AI technology in the field of cyber and network information security and social and public security.

Management

An AI security management system to effectively control AI security risks and actively promote the application of AI technology.

Security Risks Of AI

As a strategic and transformative information technology, AI has introduced new uncertainties into cyberspace security.

- **Cybersecurity risks** involve vulnerabilities in network infrastructure and learning frameworks, backdoor security issues, and systemic cybersecurity risks caused by malicious applications of AI technologies.
- **Data security risks** include training data bias in AI systems, unauthorized tampering, and security risks such as the disclosure of private data caused by AI.
- **Algorithmic security** risks correspond to algorithm design and decision-related security issues in the technical layer, as well as security risks such as black-box algorithms and algorithmic model defects.
- **Information security risks** mainly include AI technology applied to information dissemination and information content security issues for smart products and applications.
- **Societal security risks** refer to the structural unemployment brought about by the application of AI and its industrialization, which will seriously affect ethics and morality and may even cause damage to personal safety.
- **National security risks** refer to the risks to national military security and political system security brought about by risks and hidden dangers from the application of AI in military operations, public opinion, and other fields.

Security Applications of AI

AI has outstanding capabilities in data analysis, knowledge extraction, autonomous learning, intelligent decision-making, automatic control, thus AI can have many innovative applications such as:

- **Network protection** applications includes use of AI algorithms for intrusion detection, malware detection, security situational awareness, and threat early warning, etc.
- **Data management** applications refer to the use of AI technologies to achieve data protection objectives such as hierarchical classification, leak prevention, and leak traceability.
- **Information censorship** applications is the use of AI technology to assist humans in undertaking rapid review of various forms of expression and a large volume of harmful network content.
- **Smart security** applications refer to the use of AI technology to upgrade the security field from passive defence toward the intelligent direction, developing of active judgment and timely early warning.
- **Financial risk control** applications uses AI technology to improve the efficiency and accuracy of credit evaluation, risk control, etc., and assisting government departments in the regulation of financial transactions.
- **Public opinion monitoring** applications refer to the use of AI technology to strengthen national online public opinion monitoring capabilities, improve social governance capabilities, and ensure national security.

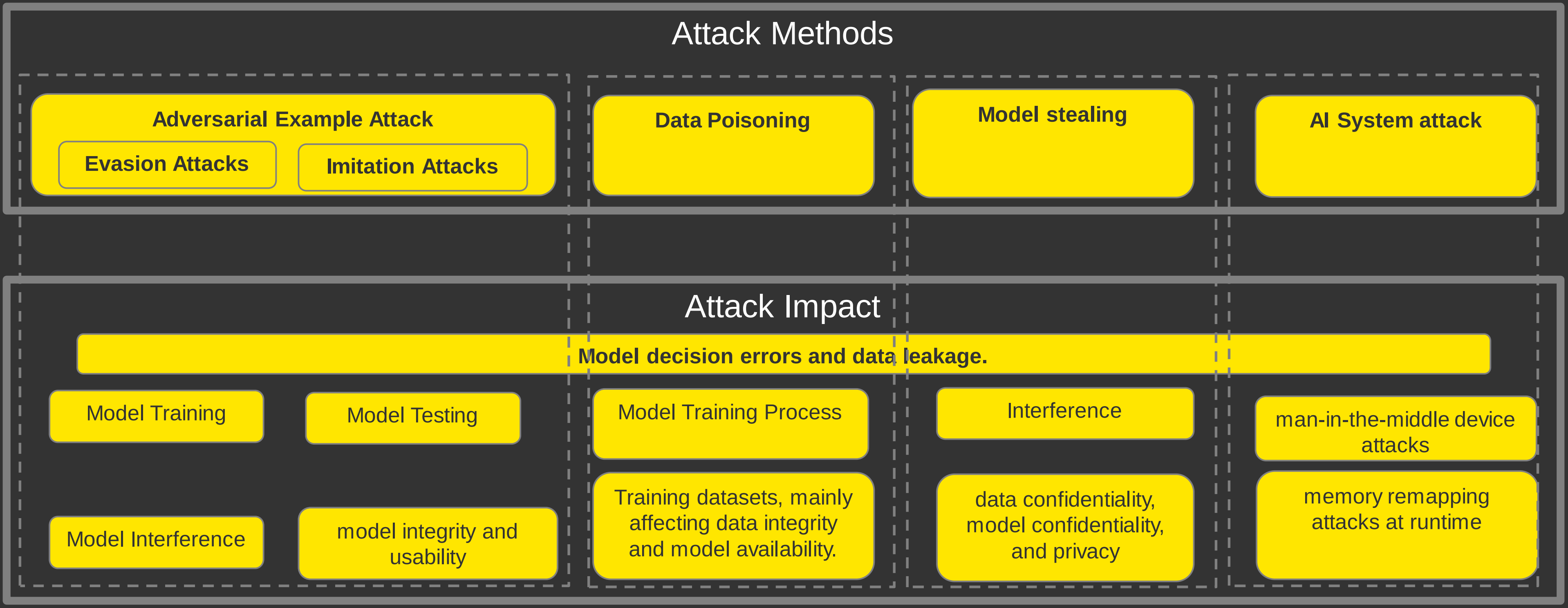
AI Security Management

Achieve effective control over AI security risks; actively promote the overall objectives for AI technology in the security domain.

- **Regulations and Policies -** establish and strengthen corresponding safety management laws and regulations and management policies for key application domains of AI and prominent security risks.
- **Standards and Specifications-** complete the formulation of international, domestic, and industry standards for AI security requirements and security assessments and evaluations.
- **Technological methods-** build technological support capabilities for security management, such as AI security risk monitoring and early warning, situational awareness, and emergency response.
- **Security assessment**, accelerate the research and development of indicators, methods, tools, and platforms for the evaluation of AI security assessments, and build third-party security assessment and evaluation capabilities.
- **Talent Development-** increase the education and training of AI talent, form a stable talent supply and an sufficient talent pool, and promote the secure and sustainable development of AI.
- **Controllable Ecology-** strengthen research and inputs at bottlenecks in the AI industrial ecology, enhance the self-guiding capability of the industrial ecology, and guarantee the secure and controllable development of AI.

Attack Methods & Impact

In addition to being threatened by traditional cyberattacks, AI also face attacks that are specific to an AI system. These attacks particularly affect systems that use machine learning



Attack Methods

Adversarial Example Attack

Evasion Attacks

Imitation Attacks

Data Poisoning

Model stealing

AI System attack

Attack Impact

Model decision errors and data leakage.

Model Training

Model Testing

Model Interference

model integrity and usability

Model Training Process

Training datasets, mainly affecting data integrity and model availability.

Interference

data confidentiality, model confidentiality, and privacy

man-in-the-middle device attacks

memory remapping attacks at runtime

Hidden Dangers of AI

Algorithm Model

An algorithm model is the core of an AI system, and security risks in the algorithm model may bring fatal security consequences to the AI system.

- ❑ Defects in robust balance and data dependence
  - Balance between accuracy and Robustness
  - Impact of datasets on accuracy
  - Reliability
- ❑ Hidden Prejudices or biases
  - Biased results or improper handling - Bias and discrimination
  - Impact of datasets on accuracy
  - Reliability
- ❑ “Black Box” - Explicability and transparency of results in AI algorithm decision
  - AI algorithms based on neural networks have “emergence” and “autonomy,”
  - Application of AI in important industries faces explicability challenges

Data Security & Privacy Protection

Data is a basic resource of AI, and machine learning requires large amounts of diverse and high-quality data for training. Each stage might pose new dangers in AI.

❑ Data Acquisition
- Excessive data acquisition
- Data acquisition that is inconsistent with user authorization.
- Compliance issues with the acquisition of personal sensitive information.
- Data quality issues.
- Difficulty in guaranteeing a user's right to opt out.

❑ Data Use
- Re-identification of anonymous data
- Data labelling and Compliance issues.
- Privacy compliance issues with automated decision-making.

❑ Data dangers at other stages
- data storage
- data sharing
- data transfers

Infrastructure

Infrastructure is software and hardware that AI products and applications generally rely on, such as software frameworks, computing facilities, and smart sensors.

- ❑ Data Acquisition
- ❑ Open Source security risks:
- ❑ Software Framework security risks
- ❑ Traditional software and hardware security risks
- ❑ System complexity and uncertainty risks
- ❑ System behaviour unpredictability
- ❑ Human-computer interaction security risks

Application

Product application such as intelligent robots and autonomous driving and Industry applications such as intelligent manufacturing, smart healthcare, and intelligent transportation

AI applications have a greater attack surface and privacy protection risks become more prominent as risks are inherited from underlying architecture will persist

- ❑ Autonomous driving: Increased network attack surfaces
  - Vulnerability risks of the physical debugging interfaces, internal microprocessors, carrying terminals operating systems, communication protocols, and cloud platforms.
- ❑ Biometric features:
  - At the data acquisition stage, AI may face attack threats such as presentation attacks, replay attacks, and illegal tampering.
  - In the biometrics storage stage, AI mainly faces threats to the biometrics database.
  - In the biometrics comparison and decision-making stage, AI faces security threats such as comparison result tampering, decision threshold tampering and hill-climbing attacks.
  - There are threats such as illegal eavesdropping, replay attacks, and man-in-the-middle attacks on biometric data transmitted between biometric feature recognition modules
- ❑ Smart speakers
  - There are vulnerabilities in the six aspects of hardware security, operating systems, application layer security, network communication security, AI security, and personal information protection.

Abuse of AI

AI can be a two sided sword - one is the improper or malicious use of AI technology to cause security threats and challenges; the second is the use of AI technology to cause uncontrollable security risks.

- ❑ Application of AI in attack methods such as fraud, dissemination of bad information, and password cracking has brought new challenges to traditional security detection.
  - Cyberattack automation has become an obvious trend
  - Spread of bad information has become more concealed
  - More and more use of AI in fraud and other illegal crimes
  - Probability of password cracking has increased
- ❑ Cross-integration of AI innovation technology into various fields promote these fields, but also the issues of AI abuse have gradually grown prominent.
  - Misuse
  - Disuse
  - Abuse

How EY Can Help?

How EY Can Help on Securing AI/ML

EY helps clients securing their AI/ML platforms end to end by offering different security services. In the era of AI and ML - well and truly here creating huge implications for business across all sectors, EY understands the heightened level of risk importance and existence of different security risks associated with AI/ML environments. EY has developed and designed different offerings to establish trust with AI/ML mitigating the security risks,

Our different services described below can help clients securing and protecting their AI/ML environments from different attack vectors -

	Secure Architecture Reviews	Threat Modelling	Security Testing	Risk Impact Assessments & Audits
Purpose	• Our secure architecture reviews assess implementations of existing AI/ML platforms end to end based on industry leading practices (such as ENISA,ETI ,IEEE etc.) in order to identify insecure practices, security configuration issues and other architecture flaws.	• Our Threat Modelling services identify the criticality of the AI/ML assets and performs a detailed threat assessment, which evaluates the different levels of asset exposure. Our TM services along with traditional security properties(CIA) also covers properties that are pertinent to the field of AI/ML such as robustness, trustworthiness, safety, transparency, accountability and data protection	• Our Security Testing services concentrate on identifying vulnerabilities and compromising issues (such as unidentified bugs, errors etc.) within the AI/ML environments at different levels (application, code, OS and Hardware). Along with the aforementioned, we also cover specific areas highlighted by the security architecture review, or threat modelling.	• Our Impact Assessments helps clients meet the legal and ethical requirements while pursuing new business goals using AI/ML. Our services focuses on all possible ethical and legal issues that can be associated with the deployment of AI/ML which partly includes the Privacy Impact Assessment (PIA), also called Data Protection Impact Assessment (DPIA) too.
Technical Process	Our process for secure architecture reviews generally covers the below domains and scope – • Assessing existing Model Training process • Data management (classification and protection) • Model integrity • Data Integrity • Resiliency • Network Protection • Algorithm security	Frameworks : • STRIDE • ENISA threat modelling methodology Core of AI/ML Threat Landscape : • Assets • Threat • Threat Actors	The process generally covers the followings – • Model Testing • Algorithm Testing • Code reviews • Penetration Testing • Vulnerability assessments • Hardware Testing	• Data Governance • Individual rights • Personal data minimization • Accountability • Transparency • Compliance • Ethical and Legal issues

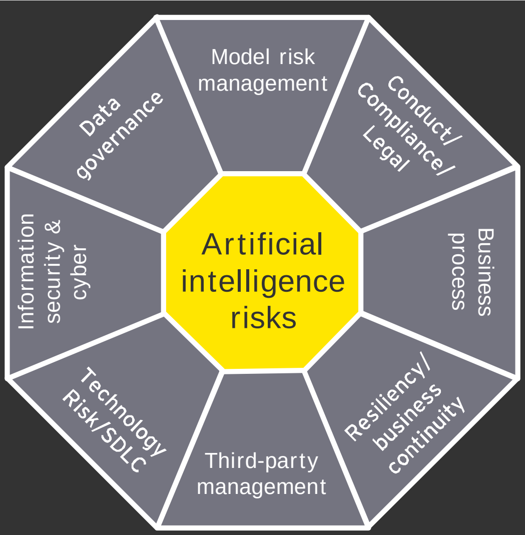
Key risks associated with AI

H	Data collection, processing and storage is not in compliance with laws and regulations
H	Data quality and completeness issues impact the accuracy of the AI outcomes

H	AI is subject to adversarial attacks which impacts its performance
H	The objective function of the AI agent is altered from its intended purpose

M	AI system is built or bought by the business without oversight from the IT team
M	AI system is put into production before it is adequately tested and appropriate monitoring is put in place

Model design, including technologies, capabilities, boundaries and training method is mismatched to intended purpose	H
Incomplete assessment of model risks due to lack of team diversity and limited focus on broad stakeholder impact	H


Artificial intelligence risks
Model risk management
Conduct/ Compliance/ Legal
Business process
Resiliency/ business continuity
Third-party management
Technology Risk/SDLC
Information security & cyber
Data governance

Lack of transparency in third-party models	M
Misplaced reliance on third-party models without sufficient independent testing and monitoring	H

AI system is not in compliance with all relevant laws and regulations	H
AI system must operate across disparate jurisdictional laws and regulations	M

AI system objective function is not aligned to business objective and intended function	H
Human operators/overseers are ill-equipped to work with the AI system	M

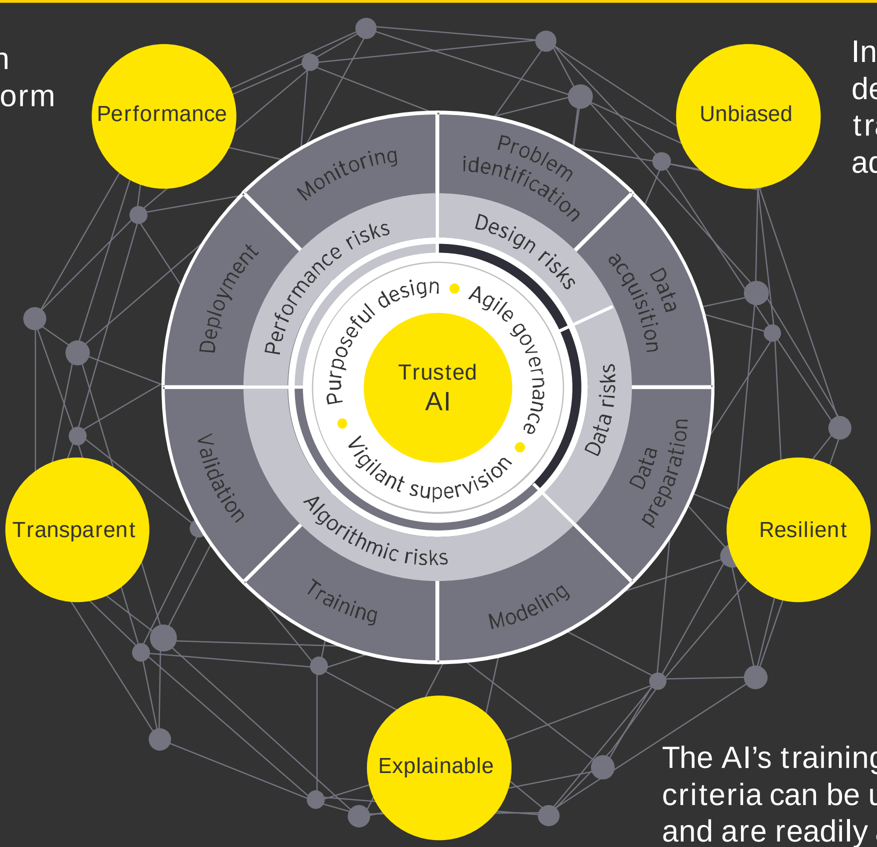
AI system performs poorly as it moves from the lab to production	M
Human operators are ill-equipped to replace AI system when inoperable	M

EY’s Trusted AI Framework

Trust in AI will require an expansion of the attributes audited

The AI’s outcomes are aligned with stakeholder expectations and perform at a desired level of precision and consistency.

When interacting with AI, an end user is given appropriate notification and an opportunity to select their level of interaction.



Inherent biases arising from the development team composition, data and training methods are identified, and addressed through the AI design.

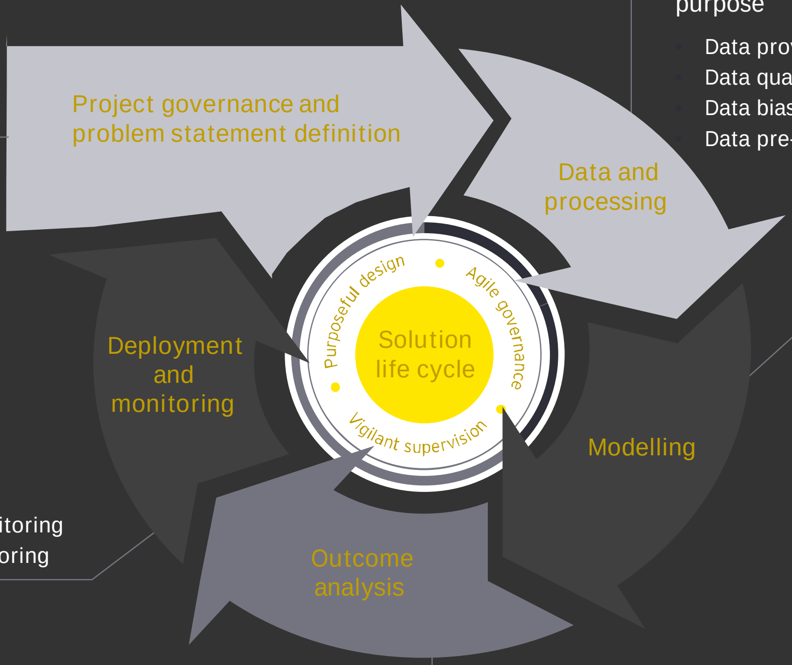
The data used by the AI system components and the algorithm itself is secured from unauthorized access, corruption and/or adversarial attack.

The AI’s training methods and decision criteria can be understood, are documented and are readily available for human operator challenge and validation.

EY’s Trusted AI Lifecycle

Business purpose, governance and stakeholder engagement are properly identified and aligned

- Business drivers
- Acceptance criteria
- Governance
- Compliance
- Reliance
- Security
- Project team



Data sourcing, profiling, processing, as well as data quality and ethical issues are lawful and fit for purpose

- Data provenance
- Data quality
- Data bias
- Data pre-processing
- Data wrangling
- Data ethics
- Pre-analysis
- Workflows

Approach and models are fit for purpose, explainable, reproducible and robust, with supporting evidence

- Model category
- Model selection
- Model build
- Feature engineering
- Reproducibility
- Bug and error handling

AI system is scalable and deployable with the right technology infrastructure, and continuously monitored

- Solution environment
- Deployment and testing
- Model management
- Workflow
- Performance monitoring
- Malpractice monitoring

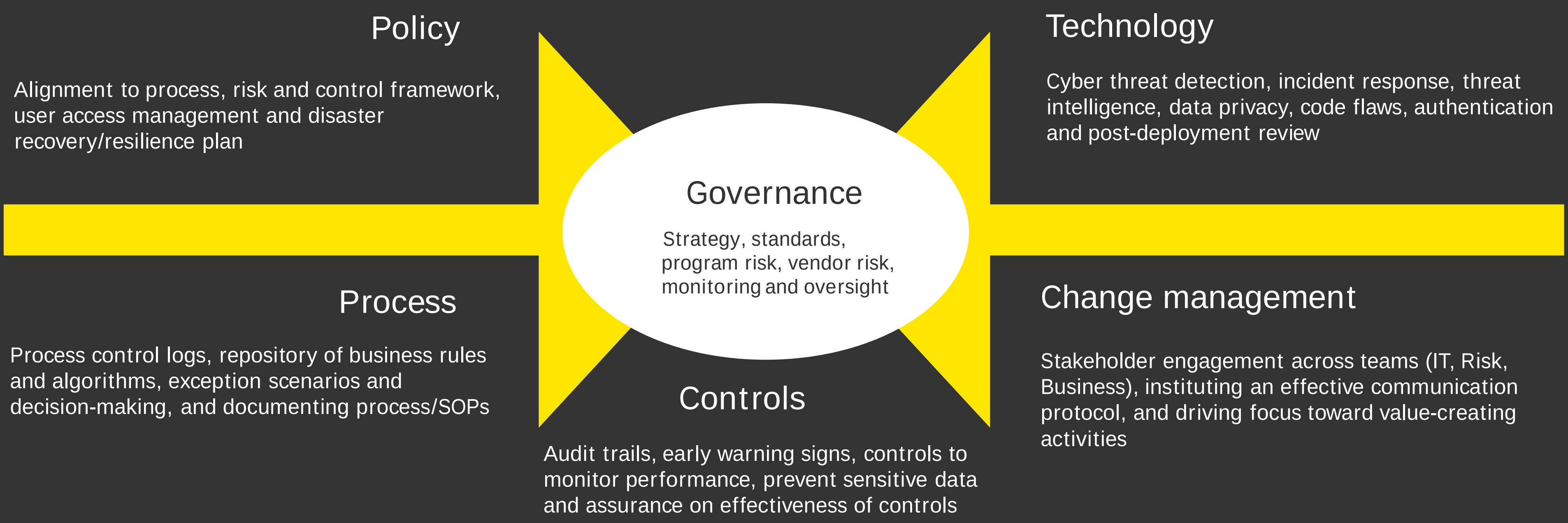
AI’s outcomes achieve desired level of precision and consistency and are aligned with ethical, lawful and fair design criteria

- Evaluation metrics
- Interpretability
- Downstream impact
- Model assessment
- Benchmarking
- Resulting actions

What’s the fix?

The cornerstones of governance, risk and control still apply

The six domains outlined below play a critical role in successful implementation of an AI program



Policy

Alignment to process, risk and control framework, user access management and disaster recovery/resilience plan

Technology

Cyber threat detection, incident response, threat intelligence, data privacy, code flaws, authentication and post-deployment review

Governance

Strategy, standards, program risk, vendor risk, monitoring and oversight

Process

Process control logs, repository of business rules and algorithms, exception scenarios and decision-making, and documenting process/SOPs

Change management

Stakeholder engagement across teams (IT, Risk, Business), instituting an effective communication protocol, and driving focus toward value-creating activities

Controls

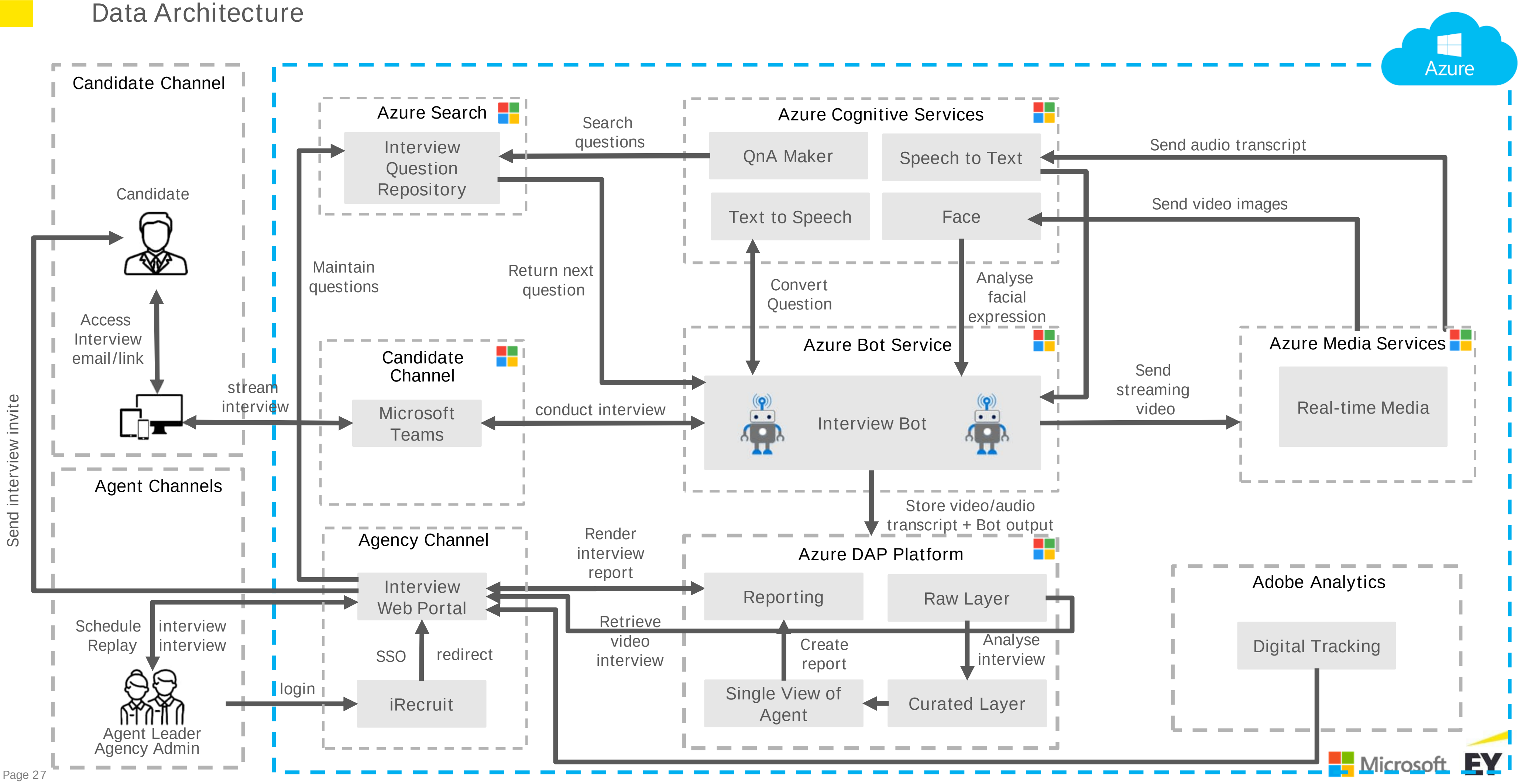
Audit trails, early warning signs, controls to monitor performance, prevent sensitive data and assurance on effectiveness of controls

Traditional risk and control categories apply to AI technology, but they each bring their own unique risk considerations

Use Cases & Case Studies

Solutions and technical requirements

Data Architecture



Insurance Company –Architectural review-Machine Learning

Client challenges

- EY was asked to perform an architecture / configuration review for a Fraud detection ML based application hosted in Azure cloud environment.
- Understanding the in scope components of the applications deployed / running in the Azure cloud.
- Examine current security control configurations for the applicable subscriptions within Client's subscribed tenancy.

The EY solution

- Evaluated current state by conducting walkthrough sessions.
- Conducted interview to understand architectural of the application and ML models.
- Various Experiments and Pipelines were reviewed to understand the flow
- Prepared a detailed questionnaire of security controls for the in scope components based on Azure Security Baselines & CIS benchmark. Checked the current configuration using this control list.
- Identified gaps and provided findings and recommendations

Project Outcome

3 main functions in the backend: data ingestion, data manipulation, model prediction and 1 user function Covered

Configuration review for:

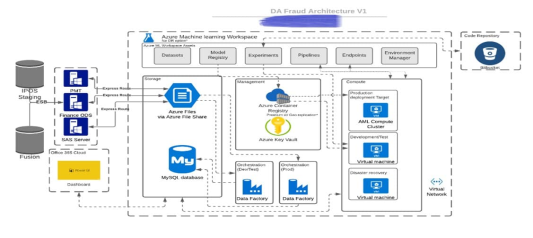
- AML Compute Cluster
- Azure Machine Learning Workspace
- Azure Data Analytics

6 Security domains

Design Validation for Machine Learning model covering

XX Experiments

XX Pipelines



Value delivered

Reduced

Security Risk and Non Compliance

Security Validated Solution

for the Machine learning model once the recommended changes are applied.

Conclusion

AI is here to revolutionize the world. AI brings in incredible opportunities, some being realized right now, others yet to come. There has been lot going on in AI R&D and companies providing AI solutions have increased at a large rate. Like any other new technology AI need specific security considerations and controls. Rather than relying on solution provider, there is a need to bring in experts to focus on specific concerns.

Human Intelligence still need to empower the AI !

Here are some questions before you move on to AI:

- Are you ready to adopt AI?
- Do you understand how AI can support your strategy?
- Do you have the resources, both human and technical, to develop and govern AI?
- Who is accountable to ensure that their AI systems are lawful, ethical and robust?
- What regulatory, financial or reputational damage could you suffer if their AI system fails? Or has this already happened?

References

https://cset.georgetown.edu/

https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-key-chinese-think-tanks-ai-security-white-paper-excerpts

https://cloud.google.com/blog/products/ai-machine-learning/build-a-transformative-ai-capability-with-ai-adoption-framework

https://towardsdatascience.com/

https://blogs.gartner.com/avivah-litan/2020/08/18/ai-security-the-dark-side-of-ai/