Bosch AIShield - Securing AI Systems

Need to Secure AI

Impact across Individuals, Organizations & Society

Individuals

- Physical Safety
  (Autonomous Vehicle Malfunction, Misdiagnose Medical Condition)
- Privacy Concerns
  (Data Breach or Acquisition without consent)
- Digital Identity Safety
  (Distortion of individual data / personal identifiable information)
- Equity & Fair Treatment
  (Racial discrimination for Underwriting & Lending)

Organizations

- Financial Performance
  (Adverse Pricing Decisions & Trading Algorithms)
- Non-Financial Performance
  (Suboptimal Algorithms for Hiring & Team/Individual Performance)
- Legal and compliance
  (Disclosure of protected consumer healthcare data)
- Reputational Integrity
  (Invasive information resulting in Advertising Claims)

Society

- National Security
  (Exposure of key military vulnerabilities/technical secrets)
- Economic Stability
  (Instability in Equity, Currency & Commodity markets)
- Political Stability
  (Manipulation of national institutional processes - elections, appointments)
- Infrastructure Integrity
  (Misuse Smart Electricity Infrastructure)

Source : McKinsey

89%
organizations did not have the right tools in place to secure their AI systems in 2021

Microsoft

60%
AI providers will include a means to mitigate possible harm as part of their AI assets & technologies by 2024

Gartner

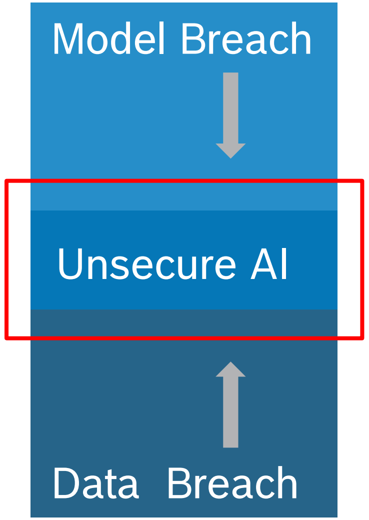
80%
is the cost difference between cyberattack scenarios, where secure AI was deployed vs not deployed

IBM Security

Top 5 Barriers to AI Implementation are associated with Security

The Case for Protection against Adversarial Threats to AI/ML Models

The Risk Mitigation for the top 5 considerations is directly co-related with how well an organization protects its AI/ML models & associated data i.e., Adversarial Threats to AI/ML Models



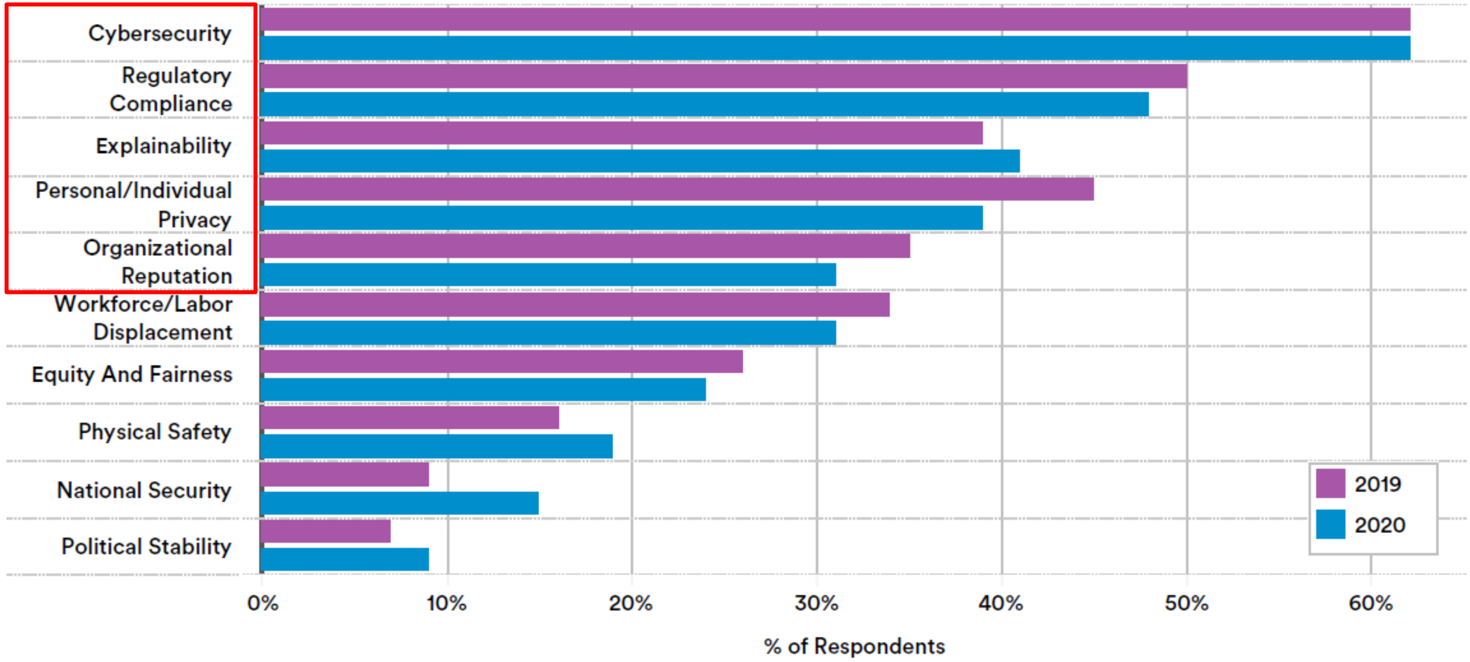
Model Breach

Unsecure AI

Data Breach

RISKS from ADOPTING AI THAT ORGANIZATIONS CONSIDER RELEVANT, 2020

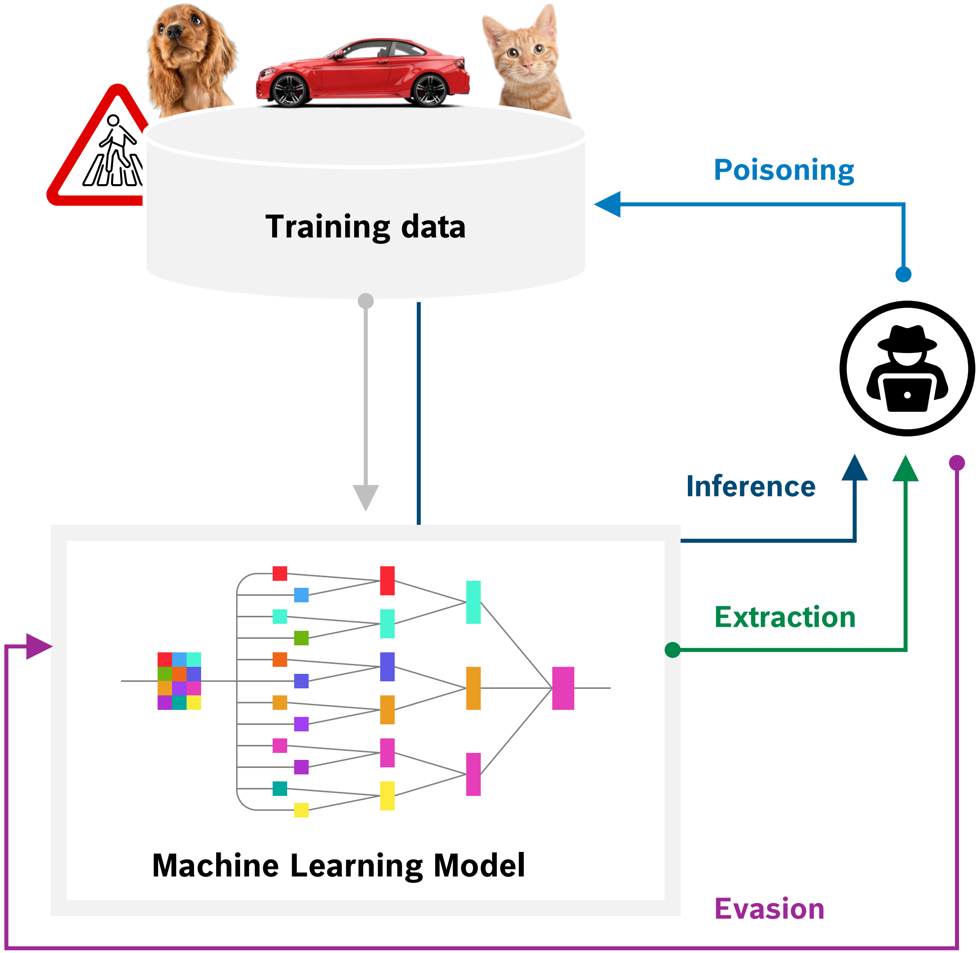
Source: McKinsey & Company, 2020 | Chart: 2021 AI Index Report



*** Gartner (2019) and McKinsey (2019,2020) found similar response from industry**

Adversarial Threats for AI/ML Models

The Genesis of Bosch AIShield



Evasion

- Spam content is embedded within an attached image to evade analysis by anti-spam models
- Threat Actors: *Cybercriminals,* Motivation: *Profit*

Poisoning

- Injecting malicious samples that subsequently disrupt the retraining process, e.g: Microsoft's Tay chatbot
- Threat Actors: *Thrill-Seekers, Hacktivists* Motivation: *Satisfaction, Ideological*

Inference

- Attempt to determine if the information of a certain record, e.g., of a person, has been part of the training data of a trained ML model or no
- Threat Actors: Cybercriminals, Hacktivist Motivation: Profit, Ideological

Extraction

- Probing ML system in order to either reconstruct the model or extract the data that it was trained on
- Threat Actors: *Insider Threats, Cybercriminals* Motivation: *Discontent, Profits*

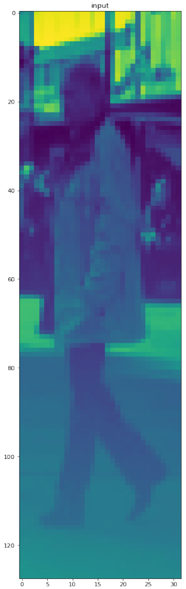
AIShield Ethical Hacking Case

Stealing/Extracting Pedestrian Detection Model for Autonomous Driving

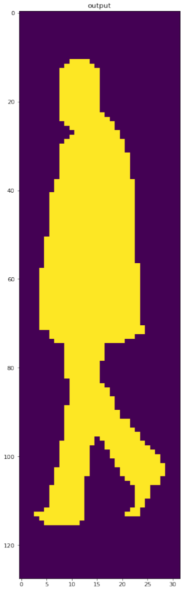
Developed over months with large proprietary datasets

~Euro 2mn *

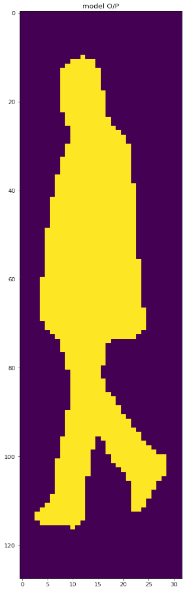
Original



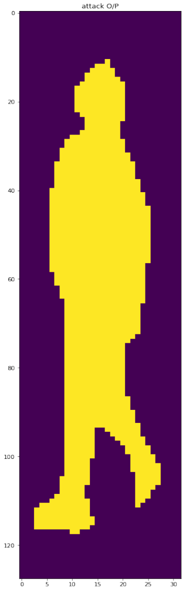
Human labeled



Original Model



Stolen Model



Stolen in <2 hours at Fraction of cost & less than 4% delta of model accuracy

Current assessment: Bosch as well as industry is not prepared fully for tackling AI Model Stealing

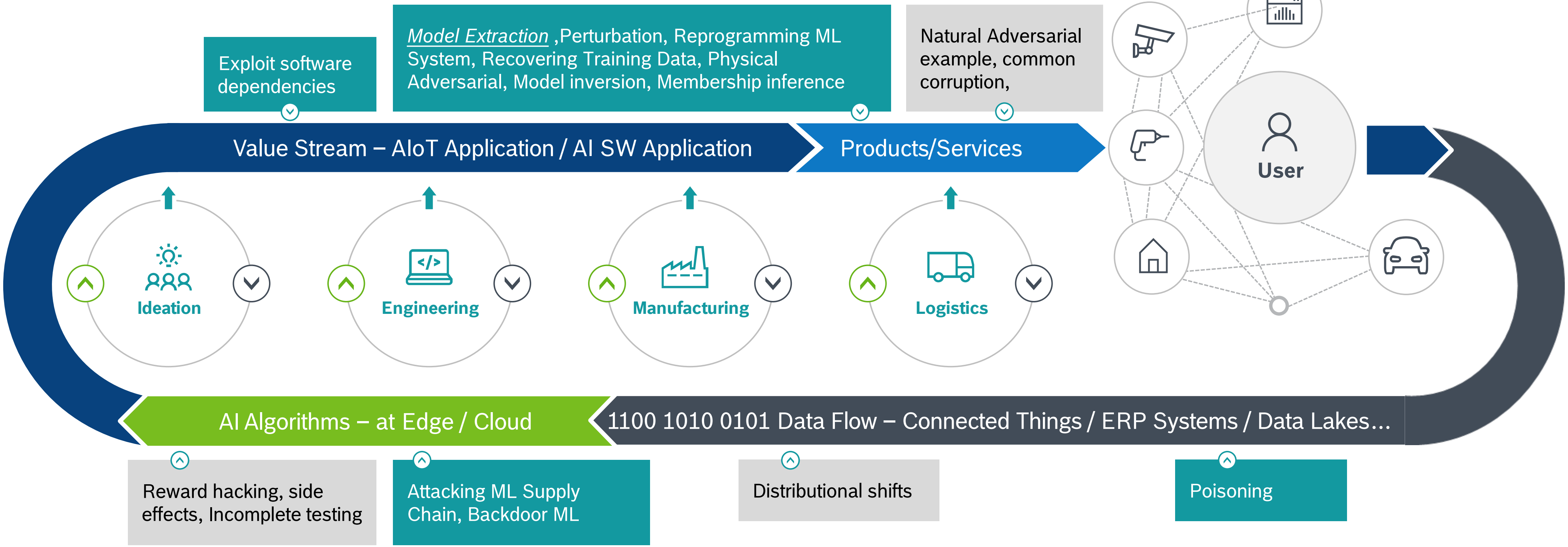
* Estimated Cost based on https://www.webfx.com/internet-marketing/ai-pricing.html; https://www.devteam.space/blog/cost-to-develop-an-ai-solution/

Security Threats across AIOT & AI S/W Lifecycle

MITRE | ATLAS

Bosch part of Consortium for Adversarial Threat Landscape for Artificial-Intelligence Systems

🟦 Value Creation 🟦 Connected Things and Services ⬛ Data 🟩 Artificial Intelligence 🟦 Intentional Failures ⬜ Unintentional Failures



Exploit software dependencies

Model Extraction ,Perturbation, Reprogramming ML System, Recovering Training Data, Physical Adversarial, Model inversion, Membership inference

Natural Adversarial example, common corruption,

Value Stream – AIoT Application / AI SW Application

Products/Services

AI Algorithms – at Edge / Cloud

1100 1010 0101 Data Flow – Connected Things / ERP Systems / Data Lakes…

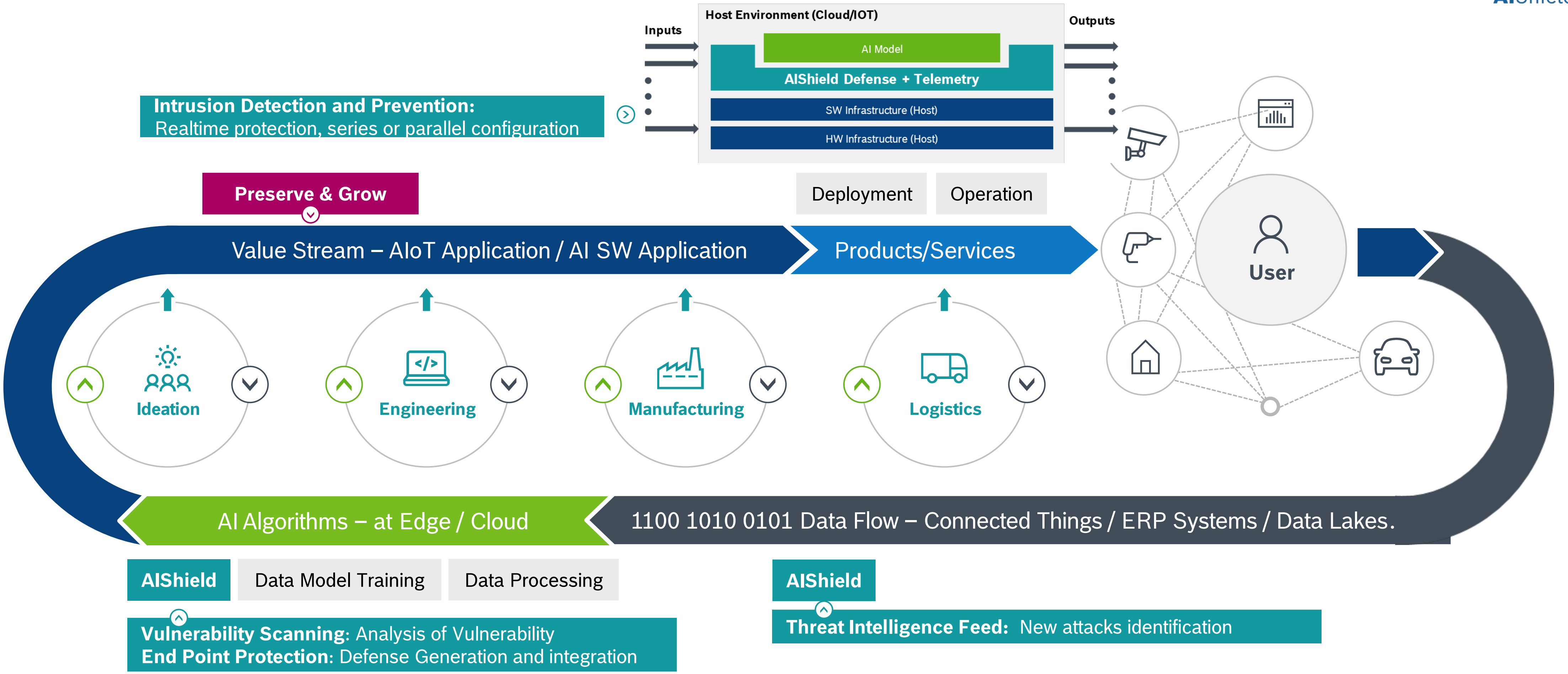
Reward hacking, side effects, Incomplete testing

Attacking ML Supply Chain, Backdoor ML

Distributional shifts

Poisoning

Security Threats' mitigation with AIShield



Host Environment (Cloud/IOT)

Inputs

AI Model

AIShield Defense + Telemetry

SW Infrastructure (Host)

HW Infrastructure (Host)

Outputs

Intrusion Detection and Prevention: Realtime protection, series or parallel configuration

Preserve & Grow

Deployment

Operation

Value Stream – AIoT Application / AI SW Application

Products/Services

User

Ideation

Engineering

Manufacturing

Logistics

AI Algorithms – at Edge / Cloud

1100 1010 0101 Data Flow – Connected Things / ERP Systems / Data Lakes.

AIShield

Data Model Training

Data Processing

Vulnerability Scanning: Analysis of Vulnerability

End Point Protection: Defense Generation and integration

AIShield

Threat Intelligence Feed: New attacks identification

Platform agnostic AI Security Solution supporting various business models

⬛ Data

🟦 Value Creation

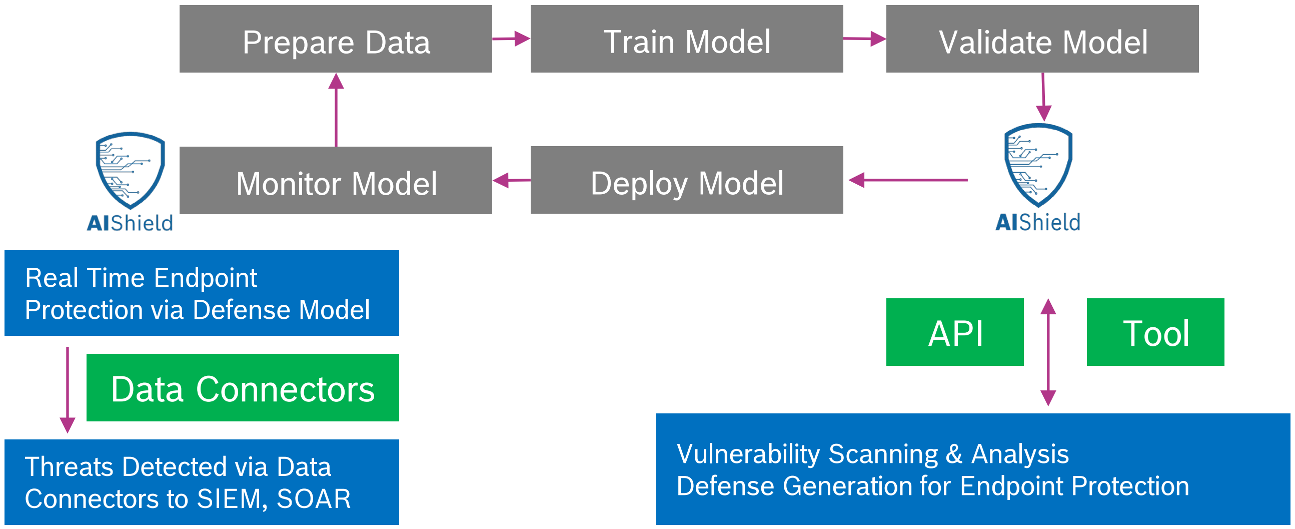
🟩 Artificial Intelligence

🟦 Connected Things and Services

Summarizing AIShield
Solution Overview

Bosch AIShield is an industry-first & patented SaaS based tool, which secures an organization’s AI assets against cyber attacks and prevents financial loss, reputational damage, loss of competitive advantage or intellectual property theft.

- Vulnerability assessment and report Generation
- Automated Defense Mechanism Generation (Patented Technology)
- Integration of Security Layer for Defense mechanism
- Suitable for Embedded and Cloud implementation
- Support for various business Models (*SaaS Tool, API SaaS for MLOps Integration*)
- Connectors for SIEM for active threat hunting and incident report triggers



Prepare Data
Train Model
Validate Model
AIShield
Deploy Model
Monitor Model
AIShield
Real Time Endpoint Protection via Defense Model
Data Connectors
Threats Detected via Data Connectors to SIEM, SOAR
API
Tool
Vulnerability Scanning & Analysis Defense Generation for Endpoint Protection

Key Benefits

- AI/ML Model Security
- Automation with MLOps
- Deployment Flexibility
- Trustworthy AI
- Brand & IP Protection

AIShield Product Offerings

AIShield	Vulnerability Scanning	Endpoint Protection / Defense Generation	Real Time Monitoring / Live Intrusion Detection & Prevention	Threat Intelligence Feed / SIEM Integration
Description	Model theft vulnerability analysis for various types of AI/ML models	Targeted defense generation and integration protecting against model extraction attacks	Real time prevention and monitoring of new attacks	Active threat hunting and incident report triggers
Functional Features	► Performs vulnerability assessment and report generation supporting >20 types of model, data type variations (e.g.: image classification, time series forecasting etc.) ► Able to ingest data, models from various storage types	► Generates targeted defense layer depending on type AI/ML of model, data type variations (e.g.: image classification, time series forecasting etc.) ► Able to integrate the generated defense with original model for plug and play operations in various configurations	► Protection against extraction attacks registered in the attack database ► Ability to protect against new attack types and register telemetry data ► Frequent attack database updates	► Report security incidents to SIEM via connectors ► Threat hunting capabilities aided by Vulnerability analysis and active monitoring ► Supports OSINT for AI Security
Usage Features	► Available as SaaS Tool, API SaaS (with MLOps Integration) ► Native support for automation		► Available as APIs and Connectors to SIEM ► Customization supported	

Thank You

Contact : Shiv Kumar
Business Head – Bosch AIShield
Kumar.shiv@bosch.com

Azure Marketplace

https://bit.ly/ampaishield

RSAC Webcast

https://bit.ly/rsacwcas

ET-CIO Article

https://bit.ly/etcioaisec