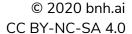
© 2020 bnh.ai CC BY-NC-SA 4.0



Phase II - Identification

	1. Assess nature and materiality of incident through:				
0	000	Disparate impact analysis Input distribution drift Residual analysis Reject on negative impact (RONI) analysis Prediction distribution drift Sensitivity analysis egorize incident as AI failure or	٥	Scanning new and past AI system traffic for: Training data Duplicate data Score insiders with affected AI system (in case of insider attack)	
Al fail	ures:		Al att	acks:	
	API m	smatches		Adversarial examples	
	Data d	rift		Deep fake	
	Data entanglement			Denial of service (DoS/DDoS)	
				Data poisoning	
				Evasion	
				Impersonation	
	Inabilit	ry to scale		Man-in-the-middle	
	Instability			Membership inference	
	Lack of accountability:			Model backdoor	
		Inability to explain		Model extraction	
		predictions		Model inversion	
		Inconsistent or inaccurate		Third-party Trojan	
		explanations		Training data breach	
		No consumer-appeal capability		Transfer learning Trojan	
	Silencing of monitoring alerts				
	System failures:				
		Application software			
		Hardware			
		Network			
	Unintended or "off-label" use				
	Unauthorized data usage				

Disclaimer: bnh.ai leverages a unique blend of legal and technical expertise to protect and advance clients' data, analytics, and Al investments. Not all firm personnel, including named partners, are authorized to practice law.





Phase II - Identification (Cont.)

3. Notify management and response staff, depending on materiality and in accordance		
with existing incident response plans:		
☐ Encrypted channels		
☐ Known impact		
☐ Need to know basis		
Agree on update/communications cadence		

Disclaimer: bnh.ai leverages a unique blend of legal and technical expertise to protect and advance clients' data, analytics, and Al investments. Not all firm personnel, including named partners, are authorized to practice law.