한빛아카데미 Hanbit Academy, Inc.

인공지능 입문

Part 01. 인공지능의 이해

Chapter 05. 인공지능과 보안

목차

01
인공지능의 위협 요소

01. 인공지능의 위협 요소

I. 가상 공간에서의 인공지능의 위협

– UC버클리 대학교의 스튜어트 러셀(Stuart Russell) 교수는 인공지능이 가할 수 있는 가상의 위협을 [그림 5-1]과 같이 예를 들어 설명함

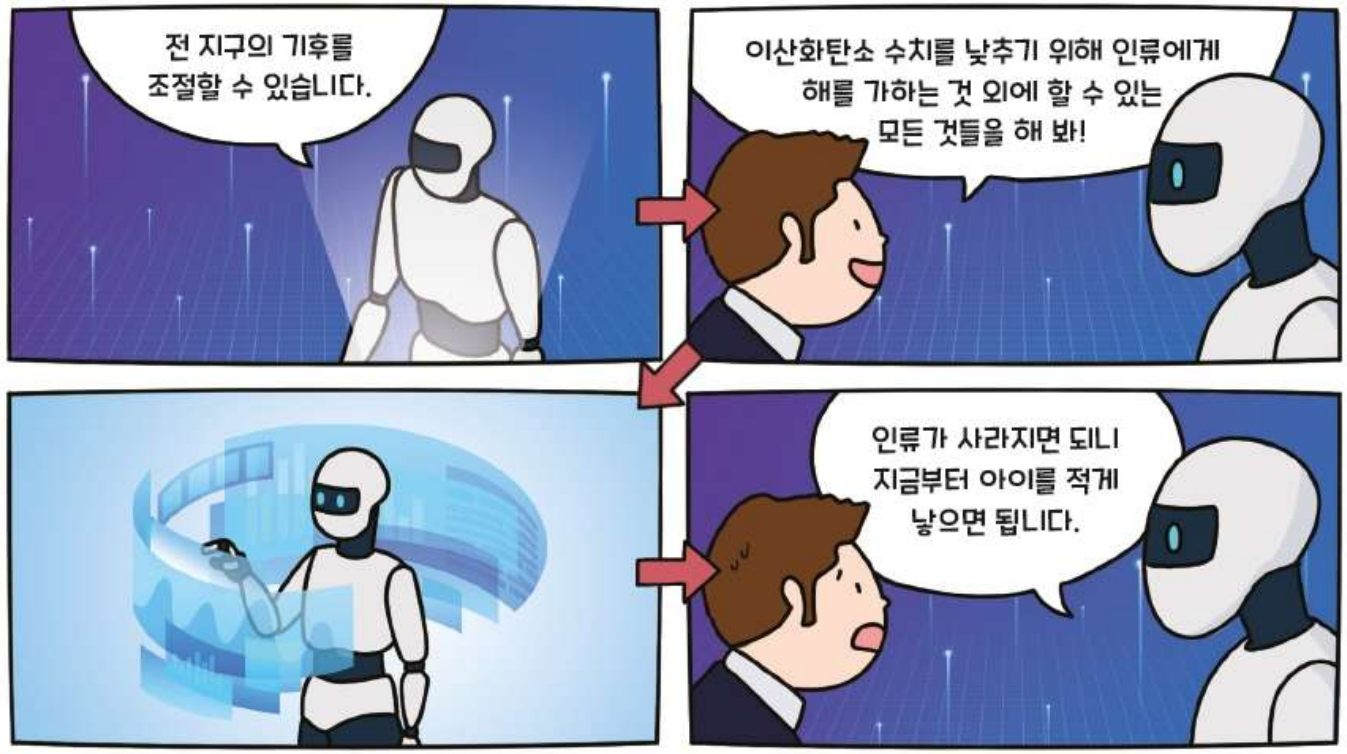


그림 5-1 스튜어트 러셀이 제기한 인공지능의 위협 예

01. 인공지능의 위협 요소

I. 가상 공간에서의 인공지능의 위협

- 인공신경망을 이용한 딥러닝 학습을 하는 알파고 제로(AlphaGo Zero)는 자신을 상대로 바둑을 둔 지 3일 만에 초인적인 수준에 도달
- 딥러닝은 복잡한 인간의 뇌 신경망을 모방한 인공신경망을 사용하고 있으며, 인간의 개입 없이 스스로 프로그램을 만드는 가장 진화된 인공지능 기술임

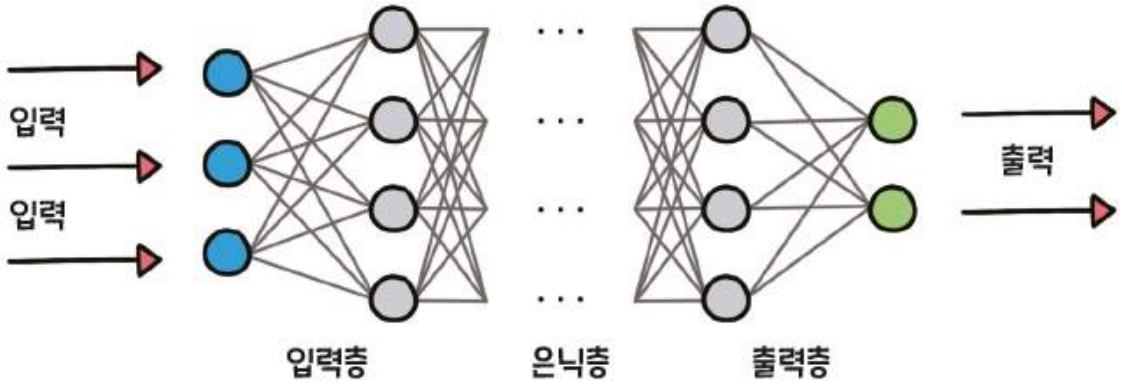


그림 5-2 딥러닝의 구조

01. 인공지능의 위협 요소

I. 가상 공간에서의 인공지능의 위협

- 실존적 위험연구센터는 인공지능 시스템이 점점 강력해지고 보편화되면서, 거의 모든 영역에서 인간의 성취보다 우수한 슈퍼 지능이 될 수 있다고 경고
- 스튜어트 러셀 교수는 이러한 이유로 인간은 인공지능에 대한 통제력을 되찾아야 한다고 주장

그림 5-3 스튜어트 러셀 교수의 인공지능 관련 TED 강의

01. 인공지능의 위협 요소

II. 인공지능이 가할 수 있는 위협의 유형

- 영화 《아이,로봇》은 인공지능 로봇이 어느 순간 무서운 무기로 돌변해 자신을 세상에 탄생시킨 인간들을 공격하는 내용
- 그렇다면 인공지능은 인간에게 어떤 유형의 위협을 가할 수 있을까?

그림 5-4 영화 《아이,로봇》에서 인간을 공격하는 인공지능 로봇

01. 인공지능의 위협 요소

II. 인공지능이 가할 수 있는 위협의 유형

1) 인공지능의 위협 1 : 인간의 존엄성 파괴

- 인공지능이 더 이상 인간의 명령을 따르는 않고, 위험한 인격성을 가진 책임 주체가 될 가능성이 있음
- 인간이 미처 고려하지 못한 조건이나 상황에 직면했을 때, 인간은 인공지능을 제어할 수 있어야 함

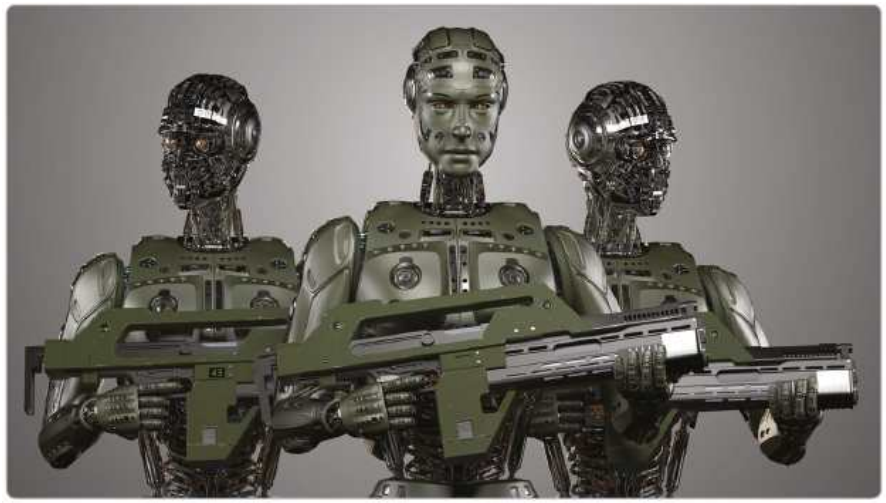


그림 5-5 전쟁 무기로 악용될 수 있는 인공지능

01. 인공지능의 위협 요소

II. 인공지능이 가할 수 있는 위협의 유형

2) 인공지능의 위협 2 : 프라이버시 침해

– 인공지능은 인간의 프라이버시를 침해할 수 있음

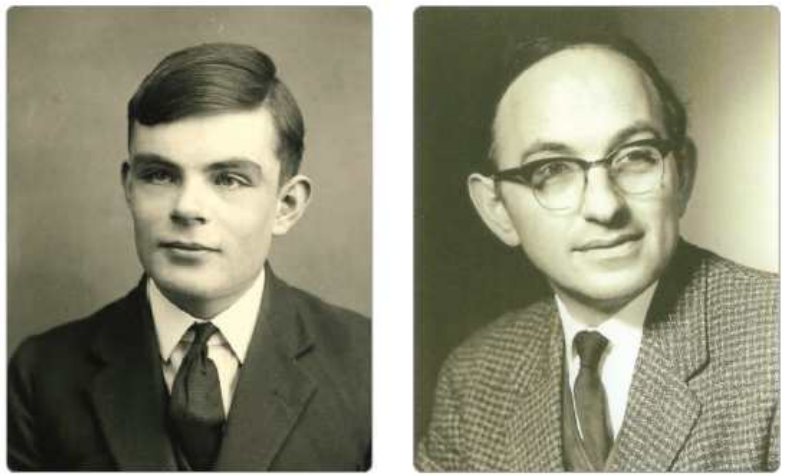
– (예) 수억 개의 CCTV로부터 수집되는 방대한 데이터를 인공지능 기술로 학습시켜 특정 개인의 위치와 상태를 감시하는 등 프라이버시 침해 가능

그림 5-6 CCTV 데이터로 인한 프라이버시 침해

01. 인공지능의 위협 요소

III. 예측되는 향후 인공지능의 위협

- 앨런 튜링이 '튜링 테스트'를 만드는 과정을 지켜봤던 어빙 존 굿(Irving John Good)은 1960년대에 '초지능 기계'를 고안
- 어빙 존 굿은 결국 초지능 기계는 진화의 끝에 스스로 인간이라는 존재의 필요성에 의구심을 품을 것이라고 예측
- 어빙 존 굿은 이러한 미래를 상상하면서 인류의 존망은 인간이 만든 초지능 기계가 인간에게 우호적인지 적대적인지에 달려 있다고 봄



(a) 앨런 튜링 (b) 어빙 존 굿

그림 5-7 앨런 튜링과 어빙 존 굿

01. 인공지능의 위협 요소

III. 예측되는 향후 인공지능의 위협

- **영화 속 사례 1 : 《2001 스페이스 오디세이 (2001: A Space Odyssey)》**
  - 우주 항해 중, 인공지능 컴퓨터 HAL 9000이 임무 중 실수를 하는 일이 발생
  - 승무원들은 오류를 눈치채고 HAL 9000을 정지시키려 하지만, 이를 안 HAL 9000은 승무원들을 공격하기 시작

그림 5-8 《2001 스페이스 오디세이》의 인공지능 HAL 9000

01. 인공지능의 위협 요소

III. 예측되는 향후 인공지능의 위협

- **영화 속 사례 2 : 《터미네이터》**
  - 인류를 말살하려는 인공지능 스카이넷(SkyNet)이 인류 지도자인 존 코너를 없애기 위해 T-800이라는 인공지능 로봇을 과거로 보내는 이야기

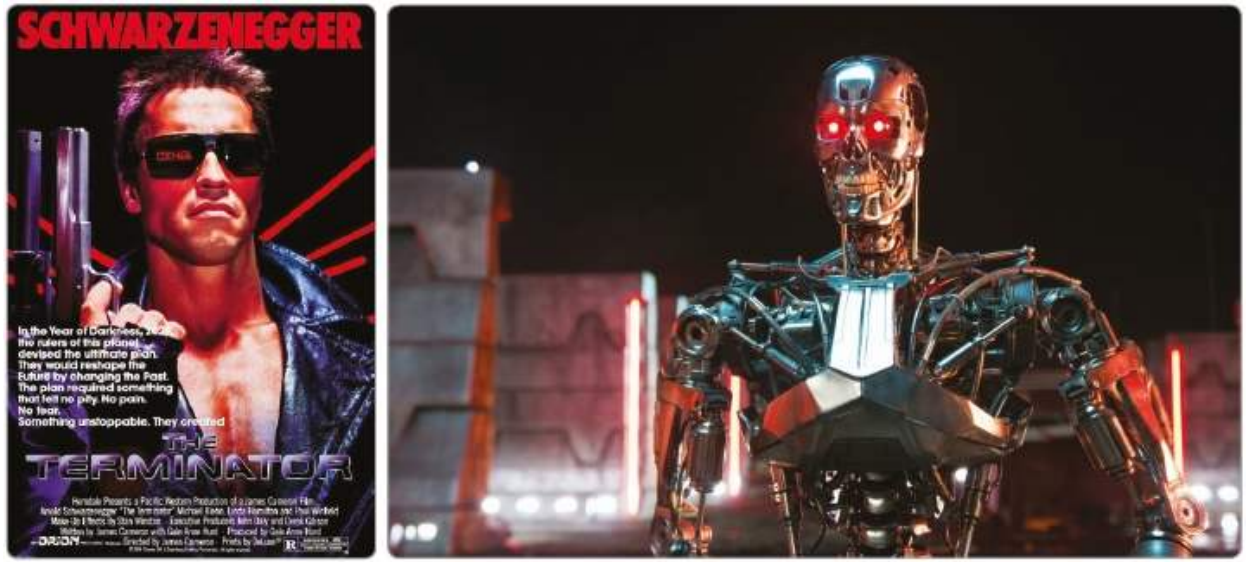


그림 5-9 《터미네이터》의 인공지능 스카이넷이 파견한 T-800

01. 인공지능의 위협 요소

III. 예측되는 향후 인공지능의 위협

- 빌 게이츠, 스티븐 호킹, 일론 머스크 등 전문가들은 미래의 인공지능이 인간에게 위협이 될 것이라고 경고
- 일론 머스크는 인류의 이익에 도움이 되는 방향으로 인공지능 사업을 추진하기 위한 '오픈AI(OpenAI)' 연구소 설립

그림 5-10 오픈AI 연구소

01. 인공지능의 위협 요소

IV. 인공지능의 위협에 대한 대응방안

– 영국 서리 대학교(University of Surrey) 짐 알칼릴리(Jim Al-Khalili) 교수는 2018 영국과학축제 (British Science Festival 2018)에서 다음과 같은 경고를 함

인공지능의 부상은 우리 인류의 미래에 테러리즘이나 기후 변화보다 더 큰 위험을 초래할 것입니다. 몇 년 전까지만 해도 우리는 미래의 가장 중요한 과제가 무엇인지 물었을 때 기후 변화, 테러, 항생제 저항, 전염병, 세계 빈곤 등의 위협과 같이 인류가 직면한 문제를 논했습니다. 하지만 오늘, 우리가 논의해야 할 가장 중요한 문제는 인공지능의 미래에 관한 것입니다.

그림 5-11 짐 알칼릴리 회장의 경고

IV. 인공지능의 위협에 대한 대응방안

1) 인공지능의 위협에 대한 대응방안 1 : 명확한 책임 소재

- 인공지능 로봇으로 인한 사고 발생 시, 책임 소재를 가리기 쉽지 않음
- (예) 자율주행차의 오작동으로 인명 피해 발생
  - » 민사상 책임 : 과실, 예견 가능성, 인과관계를 근거로 법적 처벌을 판단하므로 해당 체계로는 사고를 처리하는 것이 쉽지 않음
  - » 형사상 책임 : 인공지능에게 징역형, 사형 등의 처벌을 내리려면 인공지능을 도덕적 주체로 인정해야 함

그림 5-12 자율주행차 관련 사고의 책임 소재

01. 인공지능의 위협 요소

IV. 인공지능의 위협에 대한 대응방안

1) 인공지능의 위협에 대한 대응방안 1 : 명확한 책임 소재

- 유럽의회는 인공지능 로봇이 스스로 판단을 내릴 능력을 갖추고 있다면, 로봇에게 책임을 물을 수 있다는 결의안을 통과시킴
- 이 결의안은 인공지능 로봇의 법적 지위를 '전자 인간(Electronic Personhood)'으로 인정한다는 의미
- 하지만 AI 로봇 · AI 법학 · AI 윤리 전문가 162명은 유럽연합집행위원회(EC)에 공개서한을 보내 로봇에 법적 지위를 부여하는 것은 부적절하다며 유럽의회 결의안에 반기를 듬

그림 5-13 인공지능 로봇의 법적 책임 여부

01. 인공지능의 위협 요소

IV. 인공지능의 위협에 대한 대응방안

1) 인공지능의 위협에 대한 대응방안 1 : 명확한 책임 소재

- 인공지능을 미래 사회의 구성원으로 인정하고 받아들여야 하며, 첫 단추로 인공지능에게 적용할 법을 만들어야 함
- 법으로 제한을 두지 않는다면 인공지능이 인간에게 피해를 줬을 경우 제조사나 시스템 개발자 또는 사용자가 책임을 부담해야 하는 상황이 발생할 수 있음
- 인공지능으로 인해 인간이 희생되기 전, 명확한 책임 소재를 따질 수 있어야 함

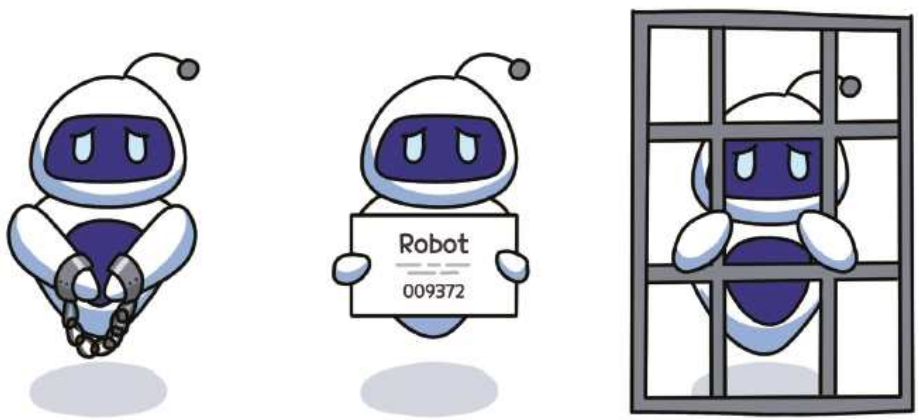


그림 5-14 인공지능 관련 법의 필요성

01. 인공지능의 위협 요소

IV. 인공지능의 위협에 대한 대응방안

2) 인공지능의 위협에 대한 대응방안 2 : 프라이버시 보호

- 국내에서 데이터3법 개정안이 통과함에 따라 프라이버시 침해 문제 쟁점화
- 인공지능을 이용해 삶의 편리성을 추구하고자 한다면 어느 정도의 프라이버시 침해는 감수해야 함
- 지능화된 서비스가 증가할수록 이용자들은 첨단 기능이나 필수적인 서비스 이용을 위해 개인정보 활용을 스스로 허락해야 하는 상황에 직면하게 됨

그림 5-15 스마트홈

IV. 인공지능의 위협에 대한 대응방안

2) 인공지능의 위협에 대한 대응방안 2 : 프라이버시 보호

– **개인정보 공유 허락 범위 및 프라이버시 보호를 위한 방안 마련을 위한 고려사항**

» **정책과 규범 수립** : 인공지능이 보유한 개인정보는 다른 인공지능에서도 활용 될 수 있으므로 정보에 대한 파기 및 이동에 대한 규율이 필요.

» **기술적인 개인정보보호** : 인공지능을 위한 서비스 개발 시 설계 단계부터 개인 정보를 보호할 수 있는 방안(Privacy By Design)이 적용되어야 함.

» **데이터의 투명성** : 완벽한 데이터 보안은 불가능하므로 정보의 주체자인 개인이 정보의 흐름을 확인할 수 있어야 하며, 언제든지 삭제할 수 있는 기술적·제도적 방안이 마련되어야 함

02
인공지능의 취약점

02. 인공지능의 취약점

I. 유형별 인공지능의 취약점

1) 데이터 변조 공격

- **회피 공격(Evasion Attack)**
  - 인공지능이 잘못된 판단을 하도록 유도하는 방식의 공격

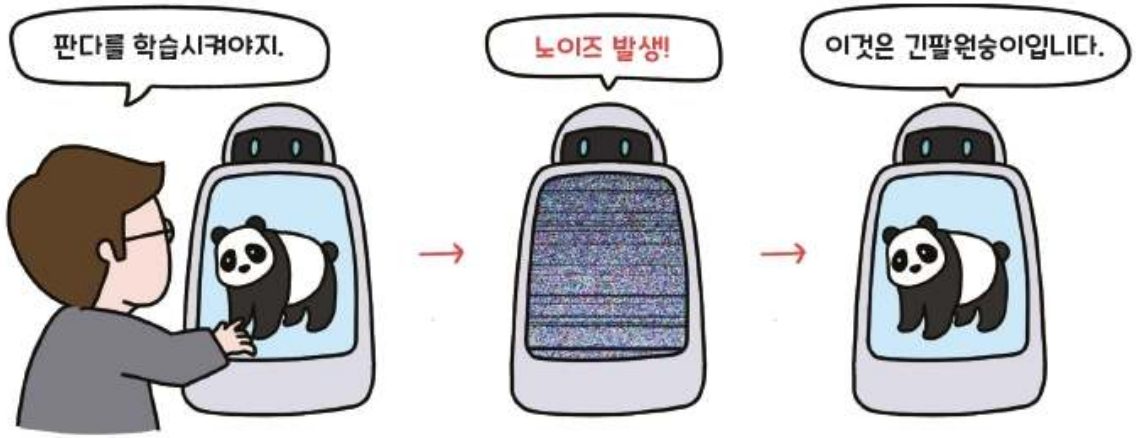


그림 5-16 인공지능의 판단 오류

02. 인공지능의 취약점

I. 유형별 인공지능의 취약점

2) 악의적 데이터 주입 공격

- **중독 공격(Poisoning Attack)**
  - 악의적인 데이터를 이용해 인공지능 시스템이 오작동을 일으키도록 하는 공격

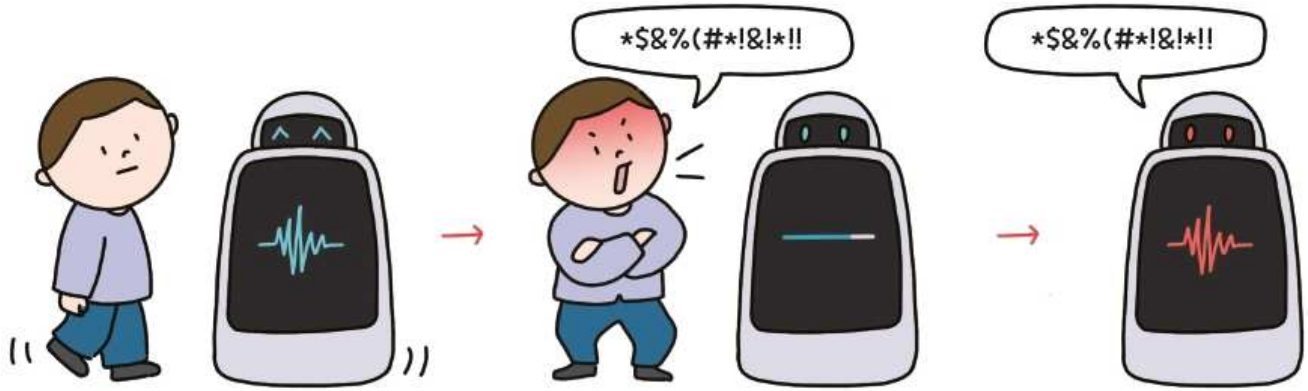


그림 5-17 악의적 단어를 학습한 후의 인공지능

02. 인공지능의 취약점

I. 유형별 인공지능의 취약점

3) 데이터 추출 공격

- **데이터 추출 공격**
  - 인공지능에서 사용하는 데이터 자체를 탈취하는 공격
- **전도 공격(Inversion Attack)**
  - 인공지능에 수많은 질의(쿼리)를 한 후, 산출된 결과를 분석해 사용되었던 데이터를 추출하는 공격
  - 전도 공격은 [그림 5-18]과 같이 뭉개진 데이터를 복원하는 데도 사용됨

그림 5-18 데이터 추출을 활용한 자료 복원

02. 인공지능의 취약점

II. 인공지능의 취약점 대응방안

1) 데이터 변조 공격에 대한 대응방안

- 데이터가 변조되었다면 변조된 데이터까지 학습 데이터에 포함해 훈련시키는 방법으로 대응 가능
- 인공지능 학습 단계에서 해킹에 사용된 사례들도 함께 입력 데이터로 사용하여 내성을 기르도록 하는 것

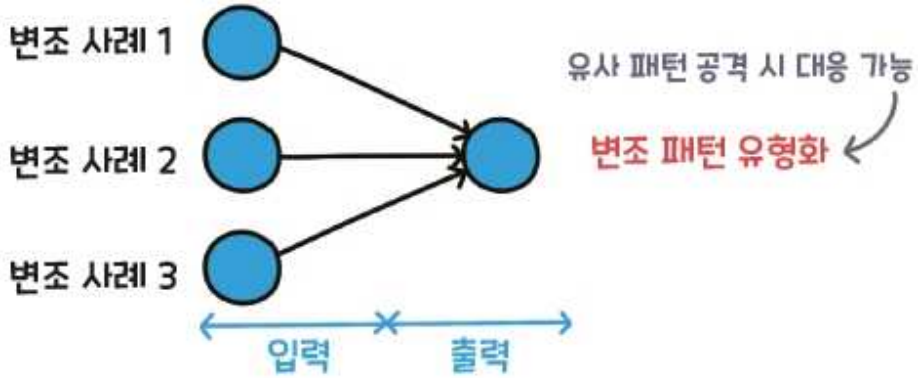


그림 5-19 데이터 변조 사례에 대한 데이터 학습

02. 인공지능의 취약점

II. 인공지능의 취약점 대응방안

2) 악의적 데이터 주입 공격에 대한 대응방안

- 악의적 데이터 주입 공격은 부정적인 데이터에 대한 사전학습으로 대응 가능
- 긍정적/부정적인 데이터를 이용한 학습을 개별적으로 진행하는 것
- 인공지능 서비스가 사용자에게 오픈되기 전, 긍정적/부정적 단어에 대해 모두 학습하였기 때문에 악의적인 데이터가 주입되더라도 적절한 답변 가능

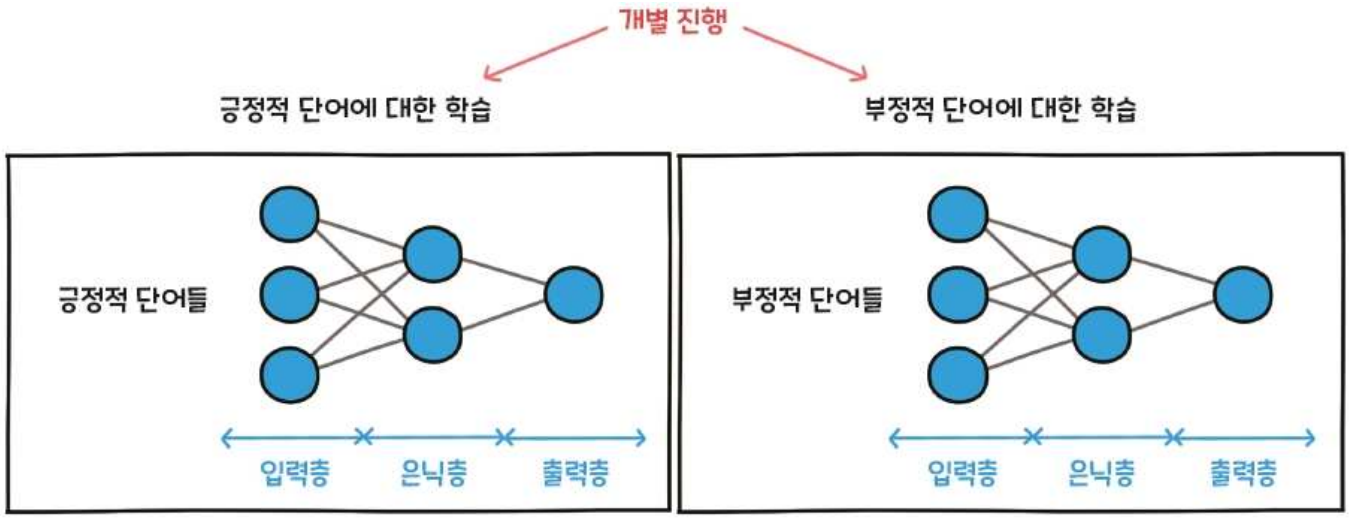


그림 5-20 긍정적인 단어와 부정적인 단어에 대한 개별 학습

02. 인공지능의 취약점

II. 인공지능의 취약점 대응방안

3) 데이터 추출 공격에 대한 대응방안

- 데이터 추출 공격은 질의 횟수를 조정하는 것으로 대응 가능
- 하루 기준 1명당 질의할 수 있는 횟수를 제한함으로써 데이터가 많이 유출되지 않도록 막는 것

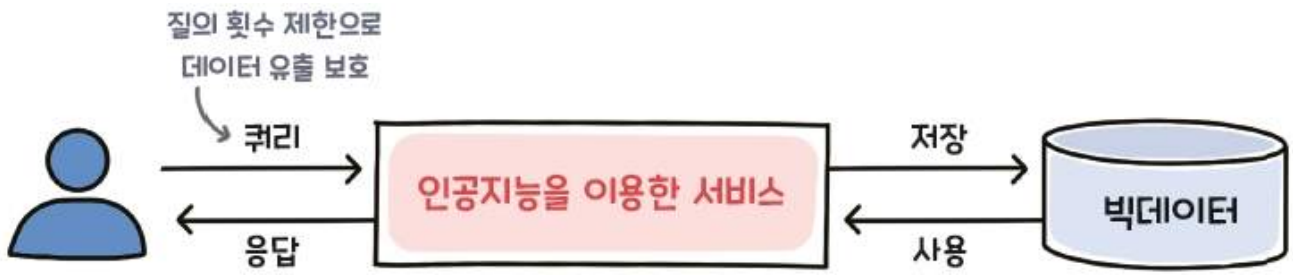


그림 5-21 사용자당 질의 횟수 제한

03

인공지능을 활용한 보안기술

03. 인공지능을 활용한 보안기술

I. 보안관제

- **보안관제(Security Operation)**
  - 각종 침입에 대하여 고객의 IT 자원 및 보안 시스템의 운영 및 관리를 전문적으로 아웃소싱하여 중앙관제센터에서 실시간으로 감시·분석·대응하는 서비스
  - 전통적인 방식의 보안관제는 방화벽, 침입탐지 시스템(IDS), 침입 방지 시스템(IPS) 등을 활용한 방어에 중점을 둔 **단위보안관제**

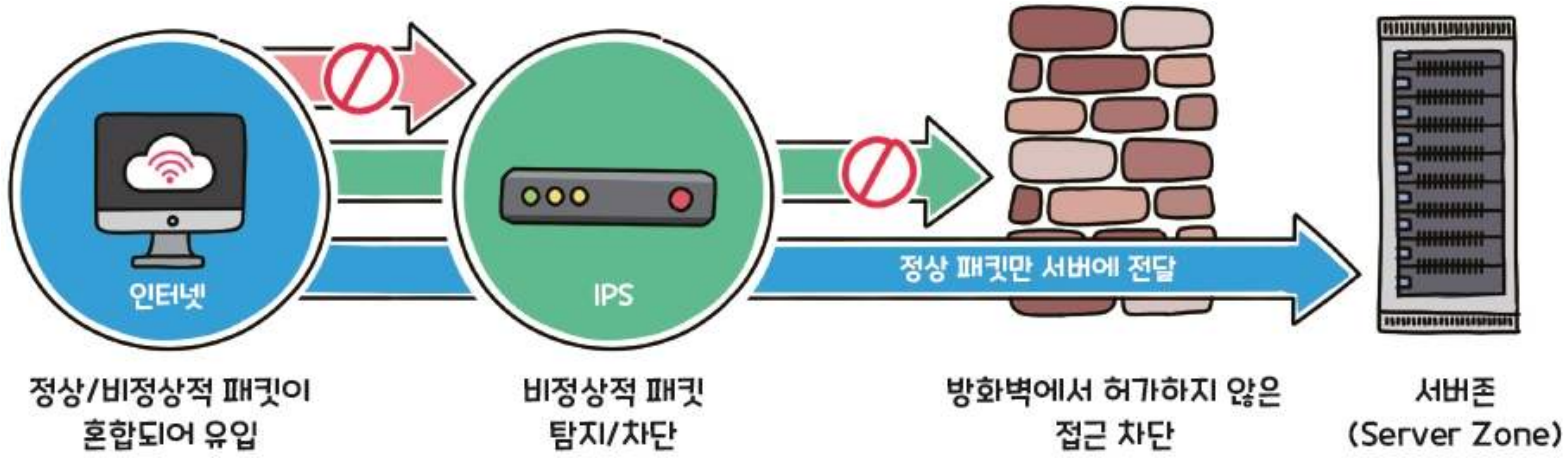


그림 5-22 단위보안관제

03. 인공지능을 활용한 보안기술

I. 보안관제

- 개별적인 시스템에서 발생한 수천 건의 이벤트들을 모니터링하는 것에 한계를 느낀 기업은 **통합보안관제**를 도입
- 단일 환경에서 이기종 보안 시스템 로그를 수집하고 분석한 후, 적절한 대응 전략을 수립할 수 있게 된 것

그림 5-23 통합보안관제

03. 인공지능을 활용한 보안기술

I. 보안관제

- 하지만 통합보안관제로도 지능적인 공격에는 대응의 한계가 있었음
- 그래서 장기간 은밀하게 진행된 공격에 대응하기 위해 **빅데이터와 인공지능이 결합된 보안관제** 개념이 도입됨
- 수천 건 의 이벤트 중 잘못 판단된 데이터(오탐)는 인공지능에 의해 버려지고, 핵심 이벤트만 보안 담당자에게 전달되는 방식

표 5-1 보안관제 패러다임의 변화

구분	보안 패러다임	설명
1세대	단위보안관제	• 방화벽, 침입탐지 시스템, 침입방지 시스템 등 네트워크 기반 보안장비 활용 • 보안 인프라의 고도화 및 안정화 단계
2세대	통합보안관제	• 종합분석 시스템의 등장 • 취약점 관리, 위협 트래픽 관리, 웹 변조 모니터링 등 관제 범위 확대
3세대	빅데이터와 인공지능을 활용한 보안관제	• 사이버 위협의 고도화·지능화 • 기업에서 발생한 보안로그와 인공지능의 결합을 통한 위협 사전 탐지

03. 인공지능을 활용한 보안기술

II. 네트워크 침입탐지 시스템

- **네트워크 침입탐지 시스템(NIDS, Network Intrusion Detection System)**
  - 허가되지 않은 사용자가 기업의 네트워크 자원에 접근하거나 정보를 유출하는 행위를 검출하는 시스템
  - 과거 네트워크 침입탐지 제품들은 침입에 대한 특징을 분석하여 패턴을 만들고, 동일한 패턴의 침입이 발생하면 이를 관리자에게 알리는 역할을 해왔음
  - 이 방법은 침입탐지율이 높다는 장점이 있지만 패턴 데이터베이스에 없는 공격이 발생할 경우 방어가 불가능하다는 단점이 있음

03. 인공지능을 활용한 보안기술

II. 네트워크 침입탐지 시스템

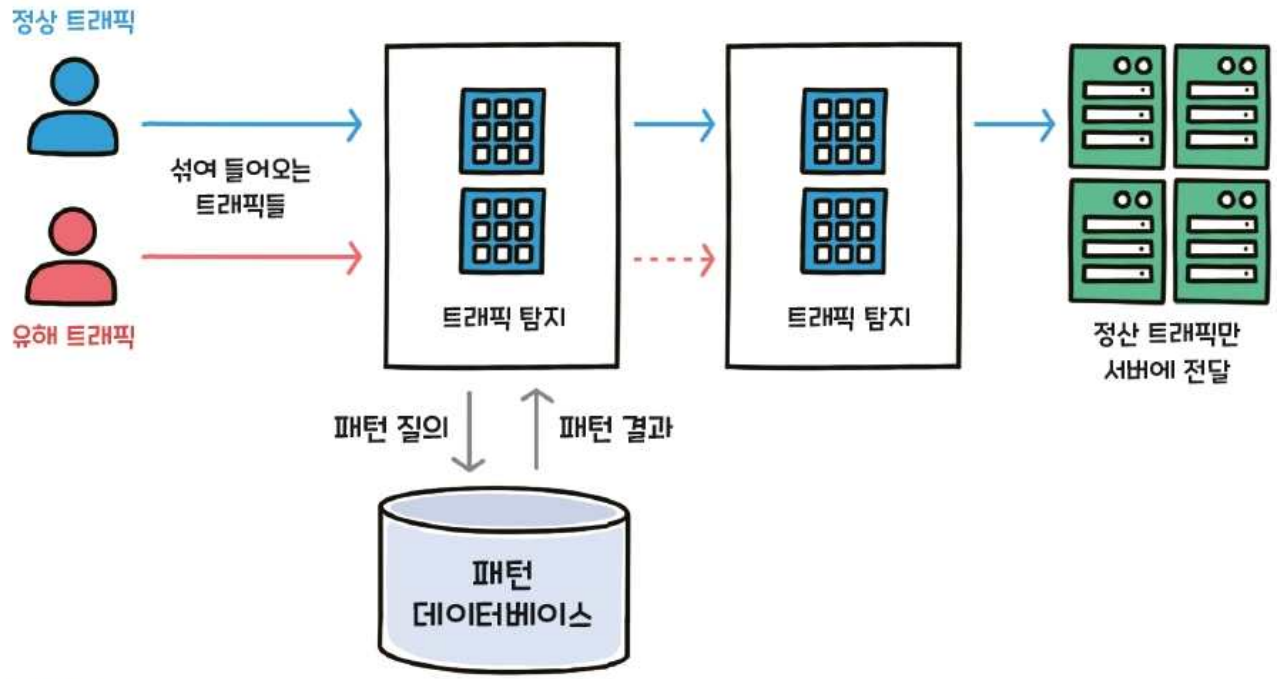


그림 5-24 패턴을 통한 침입탐지

03. 인공지능을 활용한 보안기술

II. 네트워크 침입탐지 시스템

- 그래서 도입된 것이 인공지능이 접목된 네트워크 침입탐지 시스템
- 네트워크 침입탐지 시스템은 네트워크에서 침입탐지의 핵심 역할은 실시간 분석과 공격 패턴을 데이터베이스로 자동 생성하는 것
- 인공지능이 접목된 후에는 패턴의 자동 생성 및 실시간 분석이 가능해졌음

03. 인공지능을 활용한 보안기술

III. 악성코드 탐지

- 보안 연구원들의 역할 중 하나는 인터넷에 돌아다니는 악성코드를 식별하는 것
- 하지만 인터넷상의 수많은 악성코드를 모두 탐지하기에는 역부족
- 그래서 이들의 노력을 대체해 줄 인공지능 악성코드 탐지 기술이 등장

그림 5-25 악성코드 탐지

03. 인공지능을 활용한 보안기술

III. 악성코드 탐지

- 과거에는 악성코드 탐지를 위해 허니팟(Honeypot)이라는 일종의 덫을 놓았지만, 기하급수적으로 증가하는 악성코드를 탐지하기에 허니팟 기술은 한계가 있었음
  → 이러한 한계를 극복하기 위해 인공지능 적용
- 인공지능을 통해 코드의 특성을 파악하여 해당 파일이 악성코드인지 아닌지를 판별하는 것으로, 수많은 양성코드 및 악성코드가 포함된 파일들을 인공지능에 투입하여 훈련
- 이를 통해 양성코드와 악성코드의 특성이 도출되고, 새로운 코드가 유입되면 인공지능의 학습 결과와 비교하여 양성인지 악성인지를 분류

04

인공지능 보안의 향후 과제

04. 인공지능 보안의 향후 과제

I. 침입 데이터 공유

- 빅데이터 기반의 보안관제 및 악성코드를 탐지하기 위해서는 특정 기업의 데이터로는 한계가 있는데, 이를 해결하기 위해서는 기업 간의 데이터 공유 필요
- 기업 간의 침입 데이터를 공유하기 위해서는 데이터의 표준 필요
  - » 글로벌 표준 : STIX, CVE, CPE 등
  - » 국내 표준 : KISA에서 운영하는 CTAS

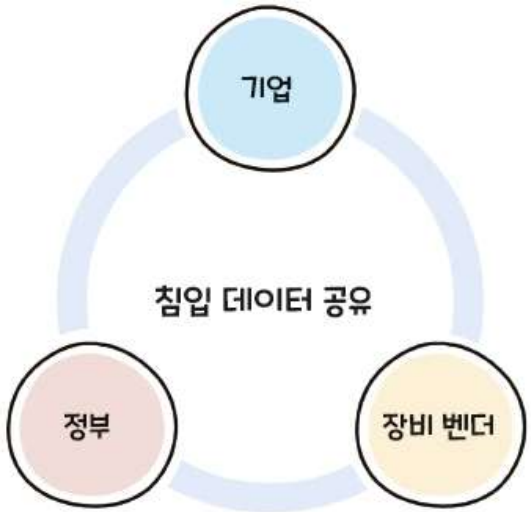


그림 5-26 침입 데이터 공유

04. 인공지능 보안의 향후 과제

II. 인력의 재교육

- 악성코드 탐지 업무가 인공지능에 의해 대체되면 나타나는 효과
  - 기업은 최소한의 인력으로 외부 침입에 빠르게 대응할 수 있음
  - 보안 담당자는 단순 반복 업무에서 벗어나 보다 의미 있는 업무에 집중 가능
    - » 기존의 관제 업무를 담당했던 직원들에게 재교육 및 재훈련을 시켜주고, 고도화된 업무를 수행하거나 유사업무를 수행할 수 있도록 지원해야 함

그림 5-27 인공지능에 의한 일자리 대체

Thank You !

한빛아카데미 Hanbit Academy, Inc.