인공지능(AI) 보안 안내서

2025. 12

과학기술정보통신부

KISA 한국인터넷진흥원

일러두기

- 본 안내서는 AI 기술 및 서비스와 관련된 보안 위협을 선제적으로 예방하고 데이터와 시스템의 안전성을 확보하며, 이를 통해 개발자, 서비스 제공자, 이용자 모두가 신뢰할 수 있는 AI 환경을 조성하는데 활용하고자 작성된 것입니다.

- 본 안내서는 과학기술정보통신부와 한국인터넷진흥원의 정책연구 사업의 연구 결과로서 내용의 무단 전재를 금합니다.

- 아울러, 안내서의 내용을 가공·인용하는 경우에는 반드시 '과학기술정보통신부·한국인터넷진흥원 《인공지능(AI) 보안 안내서》'의 출처를 밝혀 주시기 바랍니다.

- 본 안내서는 AI 서비스 및 제품을 개발하는 과정이나 서비스 제공과정에서 참고 자료로 활용할 수 있도록 편찬 되었습니다. 본 안내서 활용 시에는 기업의 업무 환경과 상황, 모델이나 시스템 개발 목적, 서비스 내용 등을 고려하여 필요하신 내용을 취사 선택하여 활용하시기 바랍니다.

- AI와 관련된 기술은 계속적으로 발전하고 있고 AI 시스템의 취약점과 이에 따른 위협 공격도 다양해지고 있기 때문에, 「인공지능(AI) 보안 안내서」는 앞으로도 지속해서 최신 기술 동향과 정보보안 위협 동향, 침해사고 사례 등을 반영해서 보안요구사항과 검증항목 내용을 업데이트해 나갈 예정입니다.

4

AI 이용자를 위한 보안 수칙

01 개요

「이용자 수칙」 개발 필요성

- AI 서비스가 안전하게 설계되었더라도, 이용자가 부주의하게 행동하면 보안 사고가 발생할 수 있다. 예를 들어, 민감한 정보를 AI 시스템에 입력하거나, 자신도 모르게 악성 AI 응용 프로그램을 실행할 경우 위험이 커질 수 있다. 해커들은 주로 이용자를 대상으로 한 사회공학적 기법(예: 피싱, 사기)을 활용하기 때문에, AI 이용자 대상 보안수칙은 이용자가 이러한 위험을 인식하고 대응하도록 하는 것이 반드시 필요하다.
- 이에, 본 안내서의 「이용자 수칙」에서는 윤리적인 원칙 선언에 그치지 않고 AI 서비스 접속·이용단계에서 **이용자가 지켜야 할 구체적인 보안 행동 지침을 제공**하고자 한다. 이용자는 AI 서비스에 입력하는 데이터가 저장되거나 악용될 가능성을 이해하지 못할 수 있다. 보안수칙에는 어떤 데이터를 입력해야 안전한지, 어떤 데이터는 공유하면 안 되는지에 대한 명확한 가이드를 제공하고자 하였다. 또한 이용자가 AI를 통해 허위 정보, AI 악용 콘텐츠, 악성 코드 등을 생성하지 않도록 가이드라인을 제공하고자 하였다.

「이용자 수칙」 도출 과정 및 참고자료

- 2024년 6월부터 구성·운영된 「AI 보안 정책 포럼」을 비롯해서 다양한 전문가 의견수렴 과정을 거쳤다. 초안 작성 후 한국인터넷진흥원(KISA)과의 협의 과정에서 몇 차례 수정작업을 거쳐 〈AI 이용자를 위한 보안 수칙〉 최종본이 마련되었다.
- 미국, 유럽, 일본 등 해외 자료를 참고하였고, 국내자료로는 국가정보원의 「챗GPT 등 생성형 AI 활용 보안 가이드라인」 등을 참고하였다.

활용 방안

- AI 서비스로 인한 피해를 사전에 예방하기 위해 AI 서비스 접속부터 이용단계까지 이용자가 지켜야 할 **구체적인 행동 지침**으로 기능할 것으로 기대한다. 또한 새로운 기술 개발이나 공격 유형이 나타날 때마다 관련 내용을 업데이트하여, 이용자 대상 교육 및 홍보 자료로도 활용 가능 할 것으로 기대된다.

02 AI 이용자에게 발생할 수 있는 보안위협 사례

본 시나리오들은 AI 서비스의 다양한 활용 사례에서 발생할 수 있는 해킹, 데이터 유출, 컨텐츠 악용 피해 등의 사례이다.

- AI 챗봇에서 중요정보 유출
  - 이용자가 입력한 이름, 주소, 금융 정보 등이 AI 챗봇 로그에 저장되고, 해커가 이를 탈취하거나, 챗봇 시스템의 취약점으로 인해 개인의 중요정보가 유출될 수 있다.
  - (사례) 챗봇이 사용자와의 대화 중에 실제 이용자들의 이름, 주소, 은행 계좌번호 등 민감한 정보를 무작위로 노출하는 사례가 발생하였다. 이는 챗봇의 학습 데이터에 포함된 중요정보가 제대로 비식별화되지 않고 사용되었기 때문으로 추정된다.
- AI 기반 음성 비서 도청
  - 스마트 AI 스피커가 음성 명령을 기다리는 동안 대화 내용을 녹음하고 이를 외부로 전송하거나 악용할 수 있다.
  - (사례 1) 아마존 에코(Amazon Echo) 사고
    - 2018년, 미국 오리건주 포틀랜드에 거주하는 한 부부의 사적 대화가 아마존의 AI 스피커 '에코'에 의해 녹음되어, 부부의 지인에게 전송되는 사건이 발생하였다. AI가 특정 단어를 호출어(wake word)로 오인하여 활성화되었고, 이후 대화를 '메시지 전송' 명령으로 잘못 인식하여 발생한 사례이다.
  - (사례 2) 구글 어시스턴트(Google Assistant) 사고
    - 2019년, 구글의 AI 음성비서 '구글 어시스턴트'에 녹음된 사용자들의 대화 1,000건 이상이 외부로 유출되는 사건이 발생하였다. 구글은 협력사 직원 중 한 명이 데이터 보안 정책을 위반하여 음성 데이터를 유출한 것으로 파악하였다.
- AI 챗봇의 악성 링크 배포
  - 해커가 AI 챗봇의 응답을 조작해 사용자에게 악성 코드가 포함된 링크를 배포한다.
  - (사례 1) 해커들이 챗GPT와 같은 AI 챗봇을 활용하여 악성 코드를 생성하는 사례가 발견되었다. 이스라엘 보안 회사 체크포인트는 챗GPT를 사용해 강력한 해킹 도구를 구축하고, 젊은 여성을 사칭해 목표물을 함정에 빠뜨리도록 설계된 새로운 챗봇을 만드는 등의 사이버 범죄 사례를 보고한 바 있다.
  - (사례 2) 피싱 이메일 작성: 챗GPT를 이용해 설득력 있는 피싱 이메일을 생성하는 사례도 증가하고 있다. 전문 지식 없이도 챗GPT를 통해 고도로 표적화된 사기 및 피싱 캠페인을 시작하는

봇 및 사이트를 구축할 수 있기 때문에 이용자들은 더욱 더 조심하고 보안 수칙에 관심을 가져야 한다.

- AI 얼굴 인식 시스템의 해킹
  - 공격자가 얼굴 인식 AI 시스템을 해킹해 다른 사용자의 권한으로 불법 접근하거나 출입을 허용할 수 있다.
  - (사례) 중국에서는 얼굴 인식 기술이 휴대전화 잠금 해제, 아파트 출입, 고속철 탑승 등 다양한 분야에 활용되고 있다. 그러나 최근 얼굴 정보를 도용하여 금융 기관에 부정 접근하거나, 딥페이크 기술을 이용해 범죄에 악용하는 사례가 늘어나고 있다. 예를 들어, 특정 지역에서는 얼굴 정보를 훔쳐 판매하려던 용의자가 체포되었으며, 이를 통해 금융 기관에 로그인하여 자금을 탈취한 사건도 발생하였다.
  - 생체 인식 정보는 유출될 경우 위·변조 등의 위험이 있으며, 한 번 유출되면 변경이 불가능하다는 특성 때문에 심각한 피해를 초래할 수 있다. 따라서 이용자들은 생체 정보를 제공할 때는 해당 AI 서비스 제공자를 더욱 더 의심하고 확인해야 한다.
- AI 서비스의 딥페이크 영상 유포
  - 공격자가 AI 기반 딥페이크 기술을 사용해 특정 사용자의 영상을 조작해 명예를 실추시키거나 사기 범죄에 활용한 사례가 있다.
  - 딥페이크를 통한 성범죄 및 명예훼손
    - (사례 1) 지인 능욕 영상 제작 및 유포: 딥페이크 기술을 이용해 특정 인물의 얼굴을 성적 영상에 합성하는 '지인 능욕' 범죄가 발생하고 있다. 이는 피해자의 사생활 침해와 명예 훼손을 초래하며, 심각한 사회적 문제로 대두되고 있다.
    - (사례 2) 청소년 대상 딥페이크 범죄: 최근 10대 청소년들을 중심으로 딥페이크 성범죄가 확산되고 있으며, 피해 학교 명단이 SNS에 떠도는 등 사회적 불안이 커지고 있다.
  - 딥페이크를 이용한 사기 범죄
    - (사례 1) 보이스피싱 및 금융 사기: 딥페이크 기술을 통해 특정 인물의 목소리나 얼굴을 합성하여 보이스피싱, 금융 사기 등 범죄에 악용되는 사례가 보고되고 있다. 예를 들어, CEO의 목소리를 모방하여 회사 재무 담당자에게 자금을 이체하도록 지시하는 등의 방식으로 실제 피해 사례가 발생하였다.
    - (사례 2) 동남아시아에서의 AI·딥페이크 활용 사기: 동남아시아 지역에서는 범죄 조직이 AI와 딥페이크 기술을 활용한 사이버 사기를 벌여 막대한 피해를 초래한 사고가 보고되었다. 유엔마약범죄사무소(UNODC)는 이러한 기술이 사칭 범죄, 딥페이크 포르노, 사이버 사기 등에 악용되고 있다고 지적하였다.

- 따라서 이용자들은 AI 서비스 접속 및 이용 시 각별히 주의해야 하고, 컨텐츠 악용에 대한 피해를 예방하기 위해 함께 노력해야 한다.

● AI를 악용한 사이버 테러 등

- ChatGPT 등 생성형 AI를 악용하여 악성코드를 생성하거나, 프로그램을 손상시키는 등 사이버 공격에 필요한 정보를 획득하거나 실제 테러를 위한 준비를 할 수 있다.
  - ▸(사례 1) 마이크로소프트(MS)의 분석에 따르면, 중국, 러시아, 북한 등은 ChatGPT를 악의적인 사이버 활동에 활용하고 있으며, 특히 북한에서는 암호화폐 탈취 등을 위해 적극적으로 생성형 AI를 이용하고 있다고 발표하였다.
  - ▸(사례 2) '25년 1월 1일 미국 라스베이거스 트럼프호텔 정문에서 발생한 테슬라의 '사이버트럭' 폭발 사건의 용의자가 ChatGPT 등 생성형 AI를 활용하였다고 경찰 당국이 발표하였으며, ChatGPT를 통해 폭발물의 목표, 특정 탄약의 이동 속도 등을 검색하고 정보를 수집하였다고 밝혔다.

AI 이용자를 위한 정보보호 수칙

01 AI 서비스 접속

1 공식 사이트에서만 다운로드
서비스 접근·가입 시,공식 사이트를 통해 가입 및 프로그램 다운로드하기

2 안전한 비밀번호 설정 및 주기적 변경
특수문자를 포함한 강력한 비밀번호로 설정하고, 주기적(3개월)으로 변경하기

3 공개된 장소에서 이용 금지
카페 등 공공장소 및 공개된 네트워크에서 서비스 이용, 정보 입력 등 금지

4 법률 및 이용약관 확인
서비스 관련 법률, 이용약관에서 정하는 금지행위, 이용자 권리 등 확인하기

02 AI 서비스 이용

1 중요 정보 입력 금지
개인정보, 기밀정보 등 중요 정보 및 허위 콘텐츠 생성을 위한 허위 정보 등 입력 금지

2 결과에 대한 정확성 검증
AI로 생성되는 정보에 대한 정확성과 함께 정보의 최신성, 편향성 등 검증 필수

3 데이터 삭제
입력, 생성된 정보는 반드시 삭제하고, 필요 시에는 별도로 다운로드하여 저장하기

4 최신 보안 업데이트 적용
서비스의 안전성, 안정성 등 유지를 위한 보안 패치는 최신버전으로 업데이트 필수

03 AI 악용 피해 예방

1 항상 의심하고 확인
정보의 허위·조작 가능성 등을 항상 의심하고 생성물 활용 분야와 목적을 반드시 확인

2 AI 악용이 의심되면 삭제
AI를 악용하여 제작된 콘텐츠, 사이트, 프로그램 등이 의심되면 삭제하고 신고하기

3 AI 악용 결과물 공유 금지
악성코드, 허위 콘텐츠 등으로 의심되는 결과물을 소장하거나 다른 사람에게 공유 금지

4 허위 콘텐츠 매매 금지
AI를 악용해 생성된 허위 콘텐츠(가짜 뉴스, 영상, 음성 등)를 돈으로 사거나 팔지 않기

03 AI 서비스 이용자를 위한 보안 수칙

01 AI 서비스 접속

1. 공식 사이트에서만 다운로드

서비스 접근·가입 시, 공식 사이트를 통해 가입 및 프로그램 다운로드 확인하기

공식 사이트나 검증된 웹 스토어를 이용해야 하는 이유

- 공식 사이트나 검증된 앱 스토어에서 제공하는 소프트웨어는 일반적으로 보안 검토를 거치기 때문에 악성 소프트웨어나 바이러스의 위험이 낮다.
- 공인된 출처에서 제공하는 AI 도구는 일반적으로 신뢰할 수 있는 개발자나 기업에 의해 만들어졌기 때문에 안정성과 성능이 보장된다.
- 공식 경로를 통해 다운로드하면 소프트웨어가 정품임을 확인할 수 있으며, 이를 통해 사용 중 발생할 수 있는 법적 문제를 피할 수 있다.
- 공식 사이트에서 다운로드한 소프트웨어는 정기적으로 업데이트가 제공되며, 문제가 발생했을 때 고객 지원을 받을 수 있는 가능성이 높다.
- 공식 플랫폼에서는 다른 사용자들의 리뷰와 평가를 통해 소프트웨어의 품질과 성능을 확인할 수 있어, 선택하는 데 도움이 된다.

이용자 주의 사항

- 사이트 접속 전에 해당 서비스의 공식 사이트인지 여부를 확인한다.
- 피싱 공격 및 거짓 사이트 방지를 위해 해당 사이트의 SSL 인증서 유무를 확인한다.
- 사용하지 않는 불필요한 AI 애플리케이션이나 확장 프로그램은 설치하지 않으며, 설치할 경우 신뢰할 수 있는 출처인지 확인한다.
- AI 서비스가 다른 서비스 또는 프로그램에 연계·확장되거나 정보가 공유되는 경우, 관련 연계 프로그램 등의 보안 취약여부 등 보안성과 안전성을 확인한다.

보안 위협

- AI 서비스 이용자에게 발생할 수 있는 보안 위협으로 가짜 사이트(Phishing 사이트 또는 Fake Website)를 통한 정보 탈취도 있다.
- 피싱 사이트는 대개 합법적인 AI 서비스 웹사이트와 매우 유사하게 만들어지며, 이용자들은 이러한 가짜 웹사이트에 자신의 로그인 정보나 신용카드 정보 등을 입력하게 된다.

표 3-1 피싱 사이트 유형

도메인 스푸핑	피싱 사이트는 합법적인 사이트의 도메인 이름과 매우 유사한 이름을 사용하여 사용자를 속임. (예) “ai-service.com” 대신 “ai-serv1ce.com”과 같이 오타를 이용한 도메인이 사용될 수 있음
합법적인 사이트 모방	사이트의 디자인, 로고, 사용자 인터페이스(UI)를 합법적인 사이트와 거의 동일하게 만들어, 이용자가 의심 없이 개인 정보를 입력하도록 유도함
이메일 피싱	이메일로 사용자를 유인하여 가짜 웹사이트로 이동하도록 만듦. 이메일에는 긴급한 메시지나 계정 문제가 있다는 내용으로 사용자의 주의를 끌어 가짜 사이트 링크를 클릭하도록 유도함

- 웹사이트 클론 공격(Website Cloning Attack)은 공격자가 AI 서비스의 웹사이트를 정확하게 복제한 가짜 사이트를 만들어 이용자를 속이는 방법이다. 이 공격은 합법적인 AI 서비스와 매우 유사하게 만들어져 있으며, 이용자에게 신뢰를 심어주어 민감한 정보를 탈취할 목적으로 사용될 수 있다.

완전한 사이트 복제	합법적인 웹사이트의 HTML, CSS, 이미지 파일 등을 그대로 복제하여 실제 사이트처럼 보이게 만듦
로그인 정보 탈취	사용자가 로그인 정보나 결제 정보를 입력하면, 이 데이터가 공격자의 서버로 전송됨

- 스피어 피싱(Spear Phishing)은 특정한 개인이나 조직을 목표로 하는 정교한 피싱 공격이다. 공격자는 특정 AI 서비스 이용자에 대한 세부 정보를 수집하고, 이 정보를 기반으로 더 신뢰할 수 있게 보이는 개인화된 가짜 웹사이트를 만들어 공격을 시도한다.

개인화된 공격	일반적인 피싱보다 더 구체적이고 개인화된 정보를 사용하여 신뢰를 높이고, 사용자가 가짜 웹사이트에 더 쉽게 속도록 만듦
조직 대상	AI 서비스를 사용하는 기업이나 단체를 대상으로 하여, 그들의 조직 내 계정이나 시스템에 접근하기 위한 수단으로 스피어 피싱이 사용될 수 있음

- 타이포스쿼팅(Typosquatting)은 이용자가 URL을 잘못 입력할 가능성을 악용하여 비슷한 도메인을 등록하는 방식으로, 이용자가 유사한 도메인 이름의 가짜 사이트에 접속하도록 유도한다. 이 공격은 이용자가 AI 서비스 웹사이트를 방문할 때 오타를 내거나 실수로 다른 사이트에 접속할 때 주로 발생한다.

가짜 사이트로 리디렉션	사용자가 잘못된 도메인에 접속하면, 가짜 로그인 페이지나 정보 탈취 페이지로 리디렉션됨

2. 안전한 비밀번호 설정 및 주기적 변경

특수문자를 포함한 강력한 비밀번호로 설정하고, 주기적(3개월)으로 변경하기

계정에 대한 보안을 강화해야 하는 이유

- 비밀번호는 사용자의 계정을 보호하는 첫 번째 방어선으로, 안전하지 않은 비밀번호를 사용하거나 주기적으로 변경하지 않으면, 해커가 계정에 접근할 위험이 커진다.
- 계정이 해킹되면 개인정보, 결제 정보, 민감한 데이터 등이 유출될 수 있다.
- 해커는 종종 무작위 대입 공격(브루트포스)이나 사전 공격을 통해 비밀번호를 알아내려고 시도한다. 강력하고 안전한 비밀번호는 이러한 공격을 방지하는 데 효과적이다.
- AI 서비스는 사용자의 행동 패턴, 검색 기록, 대화 내용 등 민감한 데이터를 포함할 가능성이 크다. 비밀번호가 안전하지 않으면 이러한 데이터를 악의적으로 이용당할 위험이 있다.
- 계정이 공격받으면 해당 서비스의 악용이나 변조가 가능해져 잘못된 정보가 생성되거나 사용될 위험이 있다.
- 계정 보안이 강화되면 해킹이나 피싱 공격으로 인한 피해를 줄일 수 있고, 강력한 비밀번호나 이중 인증을 사용하면 공격자가 계정을 접근하기 어려워진다.

이용자 주의 사항

- AI 계정 및 관련 서비스에 접근할 때에는 유추하기 어려운 강력한 비밀번호(충분한 길이 및 대소문자, 숫자, 특수 문자 혼합)를 설정해야 하고, 사이트 마다 서로 다른 비밀번호를 사용해야 한다.
- 비밀번호는 정기적으로 변경하고, 한번 사용한 비밀번호는 재사용하지 않아야 한다.

보안 위협

- AI 서비스 이용 시 비밀번호 유추 공격(Password Guessing Attack)은 해커가 사용자의 비밀번호를 추측하여 계정에 무단으로 접근하려는 공격 기법이다. 이 공격은 여러 가지 방식으로 이루어질 수 있으며, 다음과 같은 유형이 있다.
- 무작위 대입 공격(Brute Force Attack)은 해커가 가능한 모든 조합의 비밀번호를 하나씩 대입하여 맞출 때까지 시도하는 방식이다. 단순한 비밀번호나 짧은 비밀번호일수록 성공 가능성이 높다. 자동화된 프로그램을 사용하여 빠르게 많은 조합을 시도하기 때문에, 복잡한 비밀번호가 아니면 쉽

게 뚫릴 수 있다.

- 사전 공격(Dictionary Attack)은 일반적으로 사람들이 사용하는 단어나 구문(예: “password123”, “qwerty”, “123456”)을 사전에 미리 만들어 놓고, 이를 기반으로 비밀번호를 추측하는 방식이다. 사전에는 흔히 사용되는 비밀번호 목록이 포함되며, 사용자가 단순하거나 예측 가능한 비밀번호를 설정했을 경우 쉽게 뚫릴 수 있다.
- 크리덴셜 스터핑(Credential Stuffing)은 다른 웹사이트나 서비스에서 유출된 사용자 계정 정보(아이디와 비밀번호)를 사용하여 AI 서비스 계정에 접근하는 방식이다. 사용자가 여러 서비스에서 동일한 비밀번호를 사용할 경우 성공률이 높아진다.
- 사회 공학적 접근(Social Engineering)은 해커가 사용자의 개인정보(예: 생일, 이름, 애완동물 이름 등)를 조사한 후, 이를 비밀번호로 사용하는지 추측하는 방식이다. 사용자가 개인적인 정보를 비밀번호로 설정했을 경우 쉽게 유추될 수 있다.
- 스프레이 공격(Password Spraying)은 흔히 사용되는 비밀번호를 다수의 사용자 계정에 동시에 대입하여 공격하는 방식이다. 특정 계정을 집중적으로 시도하지 않기 때문에 계정 잠금 같은 방어 메커니즘을 피할 수 있다.

3. 공개된 장소에서 이용 금지

카페 등 공공장소 및 공개된 네트워크에서 서비스 이용, 정보 입력 등 금지

보안이 취약한 장소에서 이용을 자제해야 하는 이유

- 공공 Wi-Fi나 보안이 설정되지 않은 네트워크를 사용하는 경우, 해커가 네트워크를 감청하여 사용자의 로그인 정보(아이디, 비밀번호)나 데이터를 탈취할 수 있다.
- 공공 장소에서 AI 서비스에 접속하면 화면이나 입력 내용을 어깨 너머로 엿보는 것(Shoulder Surfing)이 가능하다. 특히, 비밀번호나 민감한 데이터를 입력할 때 주변 사람이 이를 볼 수 있어 정보가 노출될 위험이 있다.
- 보안이 취약한 장소에서는 사용 중인 디바이스(스마트폰, 태블릿, 노트북 등)가 도난당하거나 분실될 가능성이 높다. 디바이스가 물리적으로 탈취되면 저장된 로그인 정보, 쿠키, 인증 토큰 등이 악용될 수 있다.
- 공공 네트워크를 통해 악성 코드나 바이러스가 디바이스에 침투할 수 있다. 해커는 AI 서비스 계정을 포함한 사용자의 디지털 자산에 접근하기 위해 키로깅(Keylogging)이나 랜섬웨어를 설치할 가능성이 있다.

- 해커가 가짜 공공 Wi-Fi를 만들어 사용자가 연결하도록 유도할 수도 있다. 사용자가 가짜 네트워크에 접속하면, 해커가 모든 통신 내용을 감시하거나 데이터를 가로챌 수 있다.
- AI 서비스 사용 중 대화 내용이나 데이터 업로드 시, 서비스에 업로드되는 정보가 외부로 유출될 위험이 있다. 이러한 정보는 개인정보, 업무 기밀, 또는 기타 민감한 자료일 수 있다.
- 보안이 취약한 환경에서는 실시간으로 발생하는 보안 위협을 감지하거나 대응하기 어렵다. 보안 소프트웨어가 제대로 작동하지 않거나 업데이트되지 않은 경우, 위협에 더욱 취약해질 수 있다.

이용자 주의 사항

- 공공 Wi-Fi는 보안이 취약할 수 있으므로, 가상 사설망(VPN)을 사용해 데이터 전송을 암호화하는 것이 좋다.
- 로그인 정보나 금융 정보와 같은 민감한 데이터를 입력하는 것은 피하는 것이 안전하다.
- 웹사이트 주소가 HTTPS로 시작하는지 확인한다. 이는 데이터가 암호화되어 전송된다는 것을 의미한다.
- 사용하지 않는 네트워크 기능(예: 파일 공유, 프린터 공유 등)을 비활성화하여 보안을 강화할 수 있다.
- 사용이 끝난 후 서비스에서 로그아웃하고, 브라우저의 캐시를 삭제하는 것이 좋다.
- 공공 Wi-Fi 사용 시 비밀번호를 입력할 때는 주변 사람들에게 보이지 않도록 주의한다.
- 인터넷 연결 시 네트워크 방화벽을 활성화하고, 최신 안티바이러스 소프트웨어를 사용하여 보안을 강화한다.

보안 위협

- 맨-인-더-미들 공격(Man-in-the-Middle Attack, MITM)은 공격자가 이용자와 AI 서비스 간의 통신을 가로채어 데이터를 탈취하는 방법이다. 가짜 사이트를 사용해 사용자의 세션을 중간에서 가로챌 수도 있으며, 공격자는 사용자가 입력하는 모든 정보를 실시간으로 탈취할 수 있다.

중간에서 트래픽 탈취	공격자는 합법적인 사이트와 사용자의 통신을 가로채거나 리디렉션하여 민감한 데이터를 탈취함
네트워크 스니핑	공공 Wi-Fi나 보호되지 않은 네트워크에서 발생하기 쉬운 공격임

- 가짜 네트워크(Fake Networks) 공격은 공격자가 허위 네트워크를 통해 이용자를 속이고, 이용자가 주고받는 데이터를 탈취하는 공격이다. 이러한 공격은 공공 Wi-Fi에서 자주 발생하며, 중간자 공격(MITM)으로 확장될 수 있다.

데이터 탈취	공격자는 사용자가 AI 서비스에 로그인할 때 주고받는 로그인 정보나 기타 민감한 데이터를 가로챌 수 있음
서비스 교란	공격자가 데이터를 변조하거나 중간에서 통신을 방해하여 AI 서비스의 결과를 왜곡시킬 수 있음

4. 법률 및 이용약관 확인

서비스 관련 법률, 이용약관에서 정하는 금지행위, 이용자 권리 등 확인하기

이용약관을 확인해야 하는 이유

- 이용약관은 사용자와 서비스 제공자 간의 권리와 의무를 명확히 하므로, 이를 이해하지 않으면 발생할 수 있는 문제를 예방하기 어렵다.
- AI 서비스는 종종 사용자 데이터를 수집하므로, 이용약관에서 데이터 수집, 저장 및 사용 방식에 대한 정보를 확인해야 한다.
- 서비스 제공자의 책임이 명시되어 있는 경우, 서비스 이용 중 문제가 발생했을 때 그에 대한 법적 책임을 파악할 수 있다.
- 서비스가 종료되거나 변경될 경우의 조건과 절차에 대해 미리 알아두면, 갑작스러운 변화에 대비할 수 있다.
- 특정 산업이나 지역에서는 이용약관이 법적 요구사항을 포함하고 있을 수 있으므로, 이를 확인하여 규정을 준수할 수 있다.
- 문제가 발생했을 때의 분쟁 해결 절차나 관할 법원에 대한 정보가 포함되어 있어 상황에 따라 적절한 대응을 할 수 있다.

이용자 주의 사항

- 허용된 사용 범위: 서비스가 어떤 용도로 제공되는지, 비즈니스, 연구, 교육, 개인 용도 등 특정 목적으로 제한되어 있는지 확인한다.
- 금지된 행위: 서비스가 금지하는 활동(예: 불법적인 목적, 악의적 콘텐츠 생성, 스팸 등)을 반드시 숙지한다.
- 데이터 수집: 서비스가 어떤 데이터를 수집하고, 이를 어떻게 사용하며, 어디에 저장하는지 확인한다.
- 데이터 소유권: 사용자가 업로드한 데이터와 AI가 생성한 결과물의 소유권이 누구에게 있는지 주의 깊게 살펴본다.
- 데이터 삭제: 사용자가 데이터를 삭제하거나 계정을 해지한 경우, 데이터가 어떻게 처리되는지 확인한다.
- AI가 생성한 콘텐츠로 인해 발생한 문제(오류, 부정확한 정보, 윤리적 문제 등)에 대한 책임이 누구로 명시되어 있는지 확인한다.
- 생성 콘텐츠의 책임: AI가 생성한 결과물을 사용하는 데 있어 저작권, 윤리적 문제, 법적 문제를 방지하기 위한 이용자의 책임이 명시되어 있는지 확인한다.

- 서비스 제공자가 약관이나 정책을 변경할 경우, 사용자에게 통지하는 방법과 변경 사항에 동의하지 않을 경우의 대응 방안을 확인한다.

보안 위협

- 보안 책임의 불명확성: 이용약관을 확인하지 않으면 보안 사고 발생 시 책임이 사용자에게 귀속될 가능성을 인지하지 못할 수 있다.(예시: 이용약관에 “사용자가 입력한 데이터의 보안 책임은 사용자에게 있다“고 명시되어 있으나 이를 인지하지 못하고 데이터를 입력하여 유출 사고가 발생)
- 서비스 제공 중단 위험: 이용약관에 명시된 서비스 제공 중단 또는 데이터 삭제 조건을 알지 못하면 중요한 작업 도중 서비스가 중단되거나 데이터 접근이 불가능해질 수 있다.(예시: 무료 서비스가 갑작스레 종료되거나 유료화되면서 중요한 프로젝트가 중단)
- AI 생성물의 책임 문제: AI가 생성한 결과물(텍스트, 코드, 디자인 등)에 대해 오류나 부정확성이 있을 경우, 이에 대한 법적 책임이 사용자에게 귀속될 가능성이 있다.(예시: AI가 생성한 코드에서 보안 취약점이 발견되어 피해가 발생했으나, 약관에 “결과물의 정확성과 안전성에 대해 책임지지 않는다”고 명시된 경우)

02 AI 서비스 이용

1. 중요 정보 입력 금지

개인정보, 기밀정보 등 중요 정보 및 허위 콘텐츠 생성을 위한 허위 정보 등 입력 금지

민감하거나 중요한 정보는 입력하지 않아야 하는 이유

- AI 서비스는 데이터를 인터넷을 통해 전송하며, 이 과정에서 정보가 노출될 가능성이 있다. AI 모델이 사용하는 플랫폼의 보안 정책이 강력하더라도, 데이터 유출, 해킹, 또는 기타 예기치 않은 보안 문제가 발생할 수 있다.
- 많은 AI 서비스는 입력된 데이터를 학습 또는 개선 목적으로 저장할 수 있다. 사용자가 민감한 정보를 입력할 경우, 해당 데이터가 회사 서버에 저장되어 내부적으로 활용되거나 오용될 위험이 있다.
- AI는 입력된 데이터를 기반으로 응답하지만, 데이터를 완전히 삭제하거나 “잊어버리는” 능력이 제한적일 수 있다. 입력된 정보가 모델의 학습 데이터에 통합되거나 서비스 기록에 남아 있을 수 있다.

이용자 주의 사항

- 챗GPT 사용 시 비밀번호나 중요한 기밀사항은 절대 입력하지 않아야 한다.
- 부적절한 혹은 거짓된 정보를 입력하면 챗GPT가 그럴 듯한 오답을 생성해 허위 정보 제작 및 유포에 악용할 수 있으므로 챗GPT 등 생성형 AI를 사용할 때는 정확한 정보를 제공해야 한다.
- 사전예방을 위해 챗GPT에 질문할 수 있는 글자 수를 제한하거나 기업의 경우 사내 인트라넷에서만 챗GPT를 사용하도록 한다.
- ChatGPT(OpenAI), Copliot(MS), ClovaX(Naver) 등 AI 서비스 내 설정에서 대화 이력, 학습이력 저장(또는 전송) 기능을 비활성화하거나 데이터 저장 동의를 거부한다.

보안 위협

- 챗GPT 등 생성형 AI에 기밀 정보를 입력할 경우 해당 정보가 서비스 제공자의 직원이나 다른 위탁자에게 노출되거나, 학습 데이터로 사용될 위험이 있다.

※ 챗GPT 사용으로 인해 입력된 기밀정보가 유출된 사례

기업 내부의 기밀정보 유출	2023년 2월, 미국의 사이버보안 회사 Cyberhaven은 고객 기업에 대해 ChatGPT 사용에 관한 보고서를 발표한다. 그 보고서에 따르면, Cyberhaven 제품을 사용하는 고객 기업의 160만 명의 근로자 중, 지식 노동자의 8.2%가 직장에서 ChatGPT를 한 번이라도 사용했으며, 그 중 3.1%는 ChatGPT에 기업 기밀 데이터를 입력했다고 한다. 또한, 2023년 3월 30일, 한국의 'Economist'는 S사의 내부 일부 부서가 ChatGPT 사용을 허가한 후, 기밀 정보를 입력하는 사건이 발생했다고 보도하였다. 회사 측은 사내 정보 보안에 대한 주의를 당부하고 있었음에도 불구하고, 프로그램의 소스 코드나 회의 내용을 입력한 직원이 있었다고 발표한 바 있다.

2. 결과에 대한 정확성 검증

AI로 생성되는 정보에 대한 정확성과 함께 정보의 최신성, 편향성 등 검증 필수

AI의 결과에 확인·검증이 필요한 이유

- AI 모델은 항상 정확한 정보를 제공하지 않을 수 있다. 잘못된 정보에 의존하면 잘못된 결정을 내릴 수 있다.
- AI는 학습 데이터에 기반하여 결과를 생성하기 때문에, 데이터의 편향이 결과에 영향을 줄 수 있다. 이로 인해 왜곡된 정보가 제공될 수 있다.
- 정보는 시간이 지남에 따라 변할 수 있다. 최신 정보인지 확인하는 것이 중요하다.

- 특정 분야(예: 의학, 법률)에서는 전문적인 지식이 필요하다. AI의 정보는 참고용일 뿐, 전문가의 조언을 대체할 수 없다.
- AI 결과를 검증함으로써 정보를 보다 신뢰할 수 있게 된다. 필요 시 출처를 확인해야 한다.
- AI는 결과를 생성하는 과정에서 맥락을 고려하지 않을 수 있다. 검증을 통해 보다 깊이 있는 이해를 도모할 수 있다.

이용자 주의 사항

- AI 모델 답변이 항상 정확하거나 최신 정보를 반영하는 것은 아니므로 인터넷 검색, 전문가 의견·자문, 공식 문서 등 다양한 출처 참조를 통해 추가적으로 확인해야 한다.
- AI가 생성한 답변을 사용할 경우에는 그 출처를 표시한다.
- AI 서비스를 통해 생성된 정보는 정확성이 확보되었다고 보기 어려우므로 이용에 주의가 필요하며, 허위 정보를 입력하거나 악의적인 의도(스팸, 스미싱 등)로 사용은 금지하고, 악용 시 범죄행위임을 인식한다.

보안 위협

- 오정보 및 의사결정 오류: AI가 잘못된 정보를 제공하면 이를 기반으로 잘못된 결정을 내릴 수 있다. 예를 들어, AI가 잘못된 보안 권고를 제공하거나 취약한 설정을 권장하면 시스템이 공격에 취약해질 수 있다.
- 피싱 및 사회공학적 공격: AI를 악용해 생성된 잘못된 결과(예: 이메일 내용, 링크, 메시지)를 신뢰하면 피싱 공격에 취약해질 수 있다. 공격자가 AI를 사용하여 설득력 있는 가짜 정보를 생성하고 이를 사용자가 검증 없이 신뢰하면, 민감한 정보를 유출하거나 악성 링크를 클릭할 가능성이 높아진다.
- 취약점 악용: AI가 코드, 설정, 네트워크 구성 등에 대한 잘못된 조언을 제공하면 보안 취약점이 발생할 수 있다. 이러한 취약점은 악의적인 행위자들에게 공격 기회를 제공할 수 있다.
- 신뢰할 수 없는 소스의 정보 유입: AI는 학습 데이터에 기반하여 응답하며, 종종 공개된 인터넷 데이터나 제한된 학습 데이터에 의존한다. 이 경우, 악의적으로 조작된 데이터가 결과에 영향을 미칠 수 있다. 이를 신뢰하면 공격자들이 원하는 방향으로 시스템을 유도할 수 있다.
- 자동화된 악의적 행위: AI가 자동화된 프로세스를 지원하는 경우, 검증 없이 결과를 실행하면 악성 행위를 촉진할 수 있다. 예를 들어, AI가 추천하는 네트워크 설정을 즉시 적용하면 악의적인 코드 실행이나 백도어 생성과 같은 문제가 발생할 수 있다.

3. 데이터 삭제

입력, 생성된 정보는 반드시 삭제하고, 필요 시에는 별도로 다운로드하여 저장하기

데이터 백업이 필요한 이유

- 데이터가 저장된 채로 남아 있으면, 해커나 사이버 공격자가 이를 노릴 가능성이 있다. 특히 중요한 정보(예: 비밀 프로젝트 관련 데이터)가 포함된 경우, 보안 위협은 더욱 심각해질 수 있다.
- 저장된 데이터가 오용되거나 잘못 활용될 경우, 사용자와 서비스 제공자 모두에게 부정적인 영향을 미칠 수 있다. 특히 생성된 콘텐츠가 민감하거나 윤리적으로 논란이 될 수 있는 경우, 기록 삭제는 오용을 방지하는 데 필수적이다.
- 불필요한 데이터는 저장 공간과 처리 리소스를 차지하므로, 이를 삭제하면 서비스 운영 비용을 절감하고 효율성을 높일 수 있다.
- 필요 시에는 별도로 다운로드해 놓으면 랜섬웨어 같은 악성 공격에 대한 방어를 강화할 수 있다. 백업된 데이터가 안전하게 보관되면, 공격받더라도 손실을 최소화할 수 있다.
- 필요 시에는 별도로 다운로드해 놓으면 문제가 발생했을 때, 빠르게 데이터를 복구할 수 있다.

이용자 주의 사항

- AI 서비스 제공자의 개인정보 처리방침과 데이터 보존 정책을 검토한다.
- 서비스 제공자가 제공하는 삭제 기능이나 데이터 삭제 요청 옵션을 이용한다.
- 서비스가 데이터를 자동으로 저장하거나 클라우드에 백업하는 기능이 있다면 이를 비활성화하거나 삭제 절차를 따른다.
- 작성 도중 민감한 데이터를 입력하지 않도록 주의한다.

보안 위협

- 외부 공격: 해커나 악의적 행위자가 서버에 저장된 데이터에 접근하여 사용자 정보를 탈취할 가능성이 높아진다.
- 내부 유출: 서비스 제공자의 내부 직원이나 협력업체의 실수 또는 악의적 의도로 인해 데이터가 유출될 수 있다.
- 공격 대상 확대: 저장된 데이터가 많을수록 공격 대상이 커지고 보안 취약점이 발생할 확률이 증가한다.

4. 최신 보안 업데이트 적용

서비스의 안전성, 안정성 등 유지를 위한 보안 패치는 최신버전으로 업데이트 필수

보안 업데이트가 필요한 이유

- 소프트웨어는 시간이 지남에 따라 새로운 보안 취약점이 발견될 수 있다. 보안 패치는 이러한 취약점을 수정하여 해커가 시스템에 침투하거나 데이터를 악용하는 것을 방지한다.
- 보안 업데이트는 종종 새로운 기능이나 성능 개선도 포함되므로, 최신 버전을 유지하는 것이 전체적인 사용자 경험을 향상시킬 수 있다.
- 정기적인 보안 패치 적용은 서비스 제공자의 신뢰성을 높이며, 이용자가 서비스에 대해 안심할 수 있는 환경을 제공한다.
- AI 서비스는 종종 민감한 데이터를 다루기 때문에, 보안 패치를 통해 데이터 유출이나 손실을 방지하는 것이 중요하다.

이용자 주의 사항

- 소프트웨어 및 사용 중인 디바이스의 운영 체제, 보안 소프트웨어를 항상 최신 버전으로 업데이트한다.
- AI 서비스 및 관련 소프트웨어에서 가능한 경우 자동 업데이트를 활성화하여 보안 패치를 즉시 적용한다.
- 업데이트가 실패하거나 시스템에 문제가 생길 경우를 대비해 정기적인 데이터 백업과 복구 계획을 유지한다.
- 불필요한 AI 관련 확장 프로그램 업데이트는 AI 시스템의 성능 저하 및 보안 취약점을 초래할 수 있으므로 주의해야 한다.

보안 위협

- 취약점 악용: AI 소프트웨어의 기존 버전에는 알려진 취약점이 있을 수 있다. 최신 업데이트를 적용하지 않으면 해커가 이를 악용해 시스템에 침투하거나 데이터를 탈취할 수 있다(예: 악성 코드 실행, 권한 상승, 서비스 거부(DoS) 공격).
- 데이터 유출 및 프라이버시 침해: AI 서비스가 데이터 처리를 포함하는 경우, 보안 업데이트를 적용하지 않으면 암호화 프로토콜 또는 데이터 보호 메커니즘의 취약점이 노출될 가능성이 있다. 이러한 취약점은 민감한 데이터를 무단으로 접근하거나 외부로 유출하는 데 이용될 수 있다.
- 악성 코드 감염: 업데이트가 되지 않은 시스템은 최신 위협 탐지 및 방지 기능을 포함하지 않을 수

있다. 공격자는 이를 악용해 AI 서비스에 악성 코드를 삽입하거나 배포할 수 있다.

03 AI 악용 피해 예방

1. 항상 의심하고 확인

정보의 허위·조작 가능성 등을 항상 의심하고 생성물 활용 분야와 목적을 반드시 확인

항상 의심하고 확인이 필요한 이유

- 딥페이크와 생성형 AI의 발전
  - AI 기술이 발전하면서 딥페이크 영상, 음성 합성, 텍스트 생성이 매우 정교해지고 있고, 특히 생성형 AI를 활용하면 가짜 이미지, 영상, 기사 등을 쉽게 만들 수 있다. 이러한 허위 정보는 진짜와 구별하기 어려울 정도로 정교해 사회적 혼란을 초래할 수 있다.
  - 예시: 가짜 뉴스 영상이 유포되어 특정 정치인이나 유명인의 명예를 실추시킬 수 있다. 금융 사기에서 CEO 목소리를 합성한 딥페이크로 대규모 자금 이체를 유도한 사건도 있었다.
- 정보의 확산 속도 증가
  - AI와 인터넷의 결합으로 정보는 순식간에 전 세계로 확산될 수 있다.
  - 허위·조작된 정보는 진짜 정보보다 더 눈길을 끌기 쉬워 더 빠르게 퍼질 가능성이 높다.
  - 확인되지 않은 정보가 확산되면 개인, 기업, 국가에 막대한 피해를 줄 수 있다.
  - 예시: 잘못된 건강 정보가 퍼져 사람들이 위험한 치료법을 선택하거나 건강을 해칠 수 있고, 주가 조작을 위해 허위 기업 정보를 유포해 투자자들이 손해를 볼 수 있다.
- 개인화된 허위 정보 공격 (Targeted Manipulation)
  - AI는 사용자의 행동 패턴과 취향을 학습해 정확하게 개인화된 허위 정보를 제공할 수 있다. 이는 사용자가 허위 정보에 더 쉽게 현혹되도록 만들어 비판적 사고를 방해한다.
  - 예시: 소셜 미디어 알고리즘이 사용자가 보고 싶은 정보만 보여주며, 허위·극단적 정보의 확산을 강화한다. 사기범이 AI를 사용해 특정 개인의 정보를 조합, 맞춤형 피싱 공격을 수행한다.

이용자 주의 사항

- 정보 출처 확인
  - 공식적이고 신뢰할 수 있는 출처에서 제공된 정보인지 확인한다.
  - 뉴스, 연구 자료, 기사 등을 확인할 때 출처의 신뢰성을 검증한다.

- 소셜미디어나 커뮤니티를 통해 유포된 정보는 다시 한번 확인한다.

- 팩트 체크 및 교차 검증
  - 하나의 정보에 의존하지 말고 여러 출처를 비교하여 진위를 확인한다.
  - 팩트체크 웹사이트나 도구를 활용하여 허위 정보 여부를 검증한다.
- AI를 이용한 정보 또는 콘텐츠 등에 대한 출처 표시
  - 검증된 정보를 기반으로 AI를 이용하여 생성된 정보, 콘텐츠 등에는 반드시 AI를 이용해 생성된 정보임을 표시하여야 한다.
  - 생성된 정보, 콘텐츠에 활용된 정보의 출처를 표시한다.
  - AI를 이용하여 생성된 정보와 콘텐츠는 개인적인 용도로만 사용하여야 하며, 상업적인 목적으로 이용은 금지하여야 한다.
- 중요 정보(개인정보, 민감정보 등)가 포함되어 생성된 정보인 경우, 중요 정보에 대해 보호조치를 적용하여 활용한다.
  - 「개인정보 보호법」에 따라 개인정보와 민감정보 등을 마스킹 또는 삭제하고, 기업의 기밀정보 등에도 동일하게 적용한다.

보안 위협

- 잘못된 의사결정: AI가 제공한 정보가 조작되었거나 부정확할 경우, 이를 기반으로 한 의사결정이 잘못된 방향으로 흘러갈 수 있다.
- 사회공학적 공격(피싱 및 스캠): 공격자가 AI를 통해 설득력 있는 허위 정보를 제공하거나 사용자를 속이는 메시지를 생성할 수 있다(예시: AI가 생성한 잘못된 이메일 또는 메시지(가짜 비밀번호 재설정 요청)를 신뢰하고 실행할 경우, 계정 탈취 또는 데이터 유출 발생 가능).
- 악성 코드 및 스크립트 실행: AI가 제공한 코드나 스크립트를 검증 없이 실행하면 악성 코드가 시스템에 침투하거나 데이터를 손상시킬 수 있다(예시: AI가 제공한 “최적화된 스크립트”가 실제로는 악성 소프트웨어를 설치하도록 유도 가능).
- 데이터 손실 및 무단 접근: AI가 권장하는 설정 변경이나 데이터를 관리하는 방법이 허위 또는 악의적으로 조작된 정보일 경우, 데이터 손실 또는 무단 접근이 발생할 수 있다(예시: 잘못된 암호화 방식 또는 데이터 백업 지침을 신뢰하여 중요한 데이터를 잃거나 복구하지 못하는 상황 발생 가능).
- AI 모델 중독(Poisoning): 악의적인 사용자가 AI 서비스의 응답을 조작하여 허위 데이터를 포함시키면, 사용자는 이를 검증 없이 신뢰할 가능성이 있다(예시: AI 모델이 의도적으로 편향된 데이터로 학습된 경우, 왜곡된 결과를 제공하여 중요한 결정을 오도).

2. AI 악용이 의심되면 삭제

AI를 악용하여 제작된 콘텐츠, 사이트, 프로그램 등이 의심되면 삭제하고 신고하기

AI 악용이 의심되면 관련 콘텐츠, 프로그램 등을 반드시 삭제해야 하는 이유

- 악성 코드 및 보안 위협 방지
  - AI를 악용한 콘텐츠나 프로그램은 악성 코드, 랜섬웨어, 트로이 목마 등이 숨겨져 있을 수 있다. 이를 통해 개인정보 탈취, 시스템 감염, 데이터 손상 등 심각한 보안 사고가 발생할 수 있다.
  - 예시: 딥페이크 영상이나 AI 생성 콘텐츠에 숨겨진 악성 링크를 클릭하면, 사용자의 컴퓨터가 악성 코드에 감염될 수 있다. 또한 AI로 제작된 가짜 소프트웨어가 백그라운드에서 키로깅(타이핑 기록 감시) 등을 통해 민감 정보를 탈취할 수 있다.
- 중요 정보 유출 및 사생활 침해 예방
  - AI 악용 콘텐츠는 사용자의 이름, 얼굴, 목소리 등 개인정보를 수집하고 불법적으로 활용할 가능성이 있다. 삭제하지 않으면 해커가 이를 악용해 신원 도용, 금융 사기, 딥페이크 제작 등에 사용할 수 있다.
  - 예시: AI를 통해 유출된 목소리나 영상이 사기범에 의해 가공되어 가족이나 지인을 속이는 사기 수법에 활용될 수 있고, 악성 AI 소프트웨어가 사용자 정보를 백그라운드에서 전송할 수 있다.
- 시스템 리소스 오남용 방지
  - AI 악용 프로그램이나 콘텐츠는 종종 사용자의 시스템 리소스를 몰래 사용하여 불법 채굴(Cryptojacking)이나 봇넷의 일부로 활용될 수 있다. 이는 시스템 성능 저하, 과도한 전력 소모를 일으키고, 기기의 수명을 단축시킬 수 있다.
  - 예시: 악성 AI 프로그램이 사용자 컴퓨터를 이용해 암호화폐를 불법 채굴하거나, DDoS 공격의 일부로 활용될 수 있다.

이용자 주의 사항

- 의심스러운 파일이나 링크 실행 금지
  - 의심스러운 파일이나 링크를 절대 실행하거나 클릭하지 않아야 한다.
  - 실행된 순간 악성 코드가 설치되거나 시스템이 감염될 수 있다.
  - 특히 이메일, 메신저 등을 통해 전달된 출처 불명의 첨부 파일이나 링크는 위험하다.

- 신뢰할 수 있는 보안 소프트웨어로 검사
  - AI 악용이 의심되는 콘텐츠를 삭제하기 전, 반드시 보안 소프트웨어를 사용하여 시스템을 검사해야 한다.
  - 최신 보안 업데이트가 적용된 백신 프로그램을 사용해 전체 시스템을 점검해야 한다.

보안 위협

- 악성코드 확산 및 감염: 악의적으로 설계된 AI 기반 콘텐츠나 프로그램은 악성코드(바이러스, 랜섬웨어, 트로이 목마 등)를 포함할 수 있다. 이를 방치하면 다른 장치나 서비스로 확산될 가능성이 높다.(예시: 악성코드가 백그라운드에서 실행되어 데이터를 손상시키거나 탈취, 시스템 과부하를 초래)
- 데이터 유출: 악용 콘텐츠가 사용자 데이터를 무단으로 수집하거나 외부 서버로 전송할 수 있다.(예시: 민감한 정보(개인정보, 비밀번호, 금융 정보 등)가 외부로 유출되어 금전적 손실 및 프라이버시 침해 발생 가능)
- AI 모델 중독 및 오작동: 의심스러운 콘텐츠나 프로그램이 AI 모델의 학습 데이터를 오염시키거나 시스템의 의사결정 과정을 조작할 수 있다.(예시: AI 모델이 편향되거나 부정확한 결과를 제공하도록 유도되어 개인의 판단 오류를 초래 가능)

3. AI 악용 결과물 공유 금지

악성코드, 허위 콘텐츠 등으로 의심되는 결과물을 소장하거나 다른 사람에게 공유 금지

AI 악용 결과물은 공유를 금지해야 하는 이유

- 불법 행위에 가담할 수 있음
  - AI 악용 결과물은 불법적이거나 부적절한 목적으로 만들어졌을 가능성이 높고, 이를 공유하는 행위는 의도와 상관없이 법적 처벌 대상이 될 수 있다.
  - 딥페이크 음란물, 저작권 침해 콘텐츠, 가짜 뉴스 등을 공유하면 유포자로서 책임이 발생할 수 있다.
  - 예시: 딥페이크 성적 영상 공유 시, 「성폭력처벌법」 위반으로 징역형이나 벌금형에 처할 수 있고, 저작권 침해 콘텐츠 공유 시 「저작권법」 위반으로 과태료나 손해배상 책임이 발생할 수 있다.
- 악성 코드나 보안 위협 유포 가능성
  - AI 악용 결과물에 악성 코드, 피싱 링크 등이 숨겨져 있을 수 있고, 이를 다른 사람과 공유하면 타인의 기기나 시스템을 감염시킬 위험이 커질 수 있다.

- 예시: 악성 AI 생성 파일을 다른 사람에게 전송하면, 해당 파일을 열어보는 순간 바이러스에 감염될 수 있다. 피싱 링크가 포함된 AI 결과물을 공유해 타인의 개인정보가 유출될 가능성이 있다.

● 악용 도구의 확산 방지

- AI 악용 결과물을 공유하는 행위는 악용 도구의 확산을 조장할 수 있고, 악의적인 목적을 가진 사람들이 이를 활용해 더 많은 범죄를 저지를 수 있다.
- 예시: AI 기반으로 생성된 악성 코드나 해킹 도구가 공유되면 사이버 공격이 확산될 수 있다. 생성된 가짜 리뷰, 가짜 광고 등을 공유하면 기업이나 개인이 경제적 피해를 입을 수 있다.

이용자 주의 사항

● 의심스러운 콘텐츠 확인 및 경각심 유지

- AI를 악용해 생성된 딥페이크 영상, 가짜 뉴스, 조작된 이미지는 진짜처럼 보일 수 있으므로 항상 의심하고 확인해야 한다.
- 과도하게 자극적이거나 감정적 반응을 유도하는 콘텐츠는 즉시 공유하지 말고 확인이 필요하다.

● 링크 및 파일의 안전성 확인

- AI 악용 결과물은 악성 코드나 피싱 링크를 포함할 수 있으므로, 출처 불명의 파일, 링크는 열어보거나 다른 사람과 공유하지 않도록 주의한다.

● 콘텐츠 발견 시 즉시 신고

- AI 악용 결과물을 발견하면 해당 플랫폼이나 관련 기관에 즉시 신고해야 한다.

보안 위협

● 악성코드 및 사이버 공격 확산: AI 악용 결과물이 포함된 코드, 파일, 또는 링크가 공유되면 악성코드나 랜섬웨어가 빠르게 확산될 수 있다.(예시: 공유된 악성 AI 생성 프로그램이 사용자 장치에 감염을 일으키거나 네트워크를 통해 확산되어 광범위한 피해를 초래 가능)

● 사회공학적 공격 지원: AI를 악용해 생성된 설득력 있는 가짜 메시지(피싱 이메일, 사기 메시지, 가짜 뉴스)가 널리 퍼질 경우, 사람들이 이를 신뢰하고 민감한 정보를 제공하거나 악성 링크를 클릭하게 될 가능성이 커진다.(예시: AI 생성 피싱 이메일로 인해 대규모 데이터 유출이 발생 가능)

● 허위 정보 및 조작된 콘텐츠 확산: AI 악용 결과물로 생성된 가짜 뉴스, 허위 정보, 편집된 이미지 및 영상(딥페이크 등)이 공유되면 공공 혼란, 신뢰 상실, 정치적·사회적 불안을 야기할 수 있다.(예시: 특정 인물에 대한 딥페이크 영상이 퍼져 명예 훼손 및 신뢰성 손상 가능)

● 보안 취약점 노출: AI 악용 결과물에는 보안 취약점을 공격하는 기술적 정보나 방법론이 포함될 수 있다. 이러한 정보를 공유하면 공격자들에게 보안 취약점 악용 방법을 학습할 기회를 제공한다.(예시: 공유된 결과물이 특정 시스템의 취약점을 악용하는 스크립트인 경우, 해당 시스템이 대규모로

공격당할 위험 증가)

- 범죄 및 불법 활동 지원: AI 악용 결과물이 범죄 행위(사기, 해킹, 테러 등)에 사용될 수 있는 도구(예: 가짜 문서 생성, 스팸 메시지 자동화)를 포함하고 있다면 이를 공유함으로써 불법 활동이 확산될 가능성이 있다.(예시: AI 생성 허위 신분증이나 금융 서류가 범죄 조직에 의해 악용 가능)
- 신뢰 손상: AI 악용 결과물을 공유한 사람이 조직이나 기업의 일원일 경우, 해당 조직의 신뢰성과 명성이 훼손될 수 있다.(예시: 조직 내 직원이 AI 악용 결과물을 공유하여 외부에서 법적·윤리적 문제로 비난받음)

4. 허위 콘텐츠 매매 금지

AI를 악용해 생성된 허위 콘텐츠(가짜 뉴스, 영상, 음성 등)를 돈으로 사거나 팔지 않기

AI를 악용해 생성된 허위 콘텐츠는 매매나 공유를 금지해야 하는 이유

- 불법 행위 해당
  - 허위 콘텐츠의 생성, 공유, 매매는 의도와 관계없이 불법 행위로 간주될 수 있다.
  - 특히 AI를 악용해 생성된 콘텐츠가 허위사실 유포, 명예훼손, 저작권 침해, 성범죄와 연관된 경우, 법적 처벌을 받을 수 있다.
- 피해 확산과 2차 가해 유발
  - AI로 생성된 허위 콘텐츠(딥페이크 영상, 조작된 뉴스, 합성 이미지)는 피해자에게 큰 정신적·사회적 고통을 준다. 이를 공유하거나 매매하면 피해 확산과 함께 2차 가해를 유발하게 된다.
  - 예시: 딥페이크 음란물이나 허위사실을 퍼뜨리면 피해자의 사회적 평판이 실추되고 고통이 가중된다. 콘텐츠가 온라인에 영구적으로 남아 피해자가 일상생활에 어려움을 겪게 된다.
- 허위 정보의 빠른 확산으로 사회 혼란 초래
  - AI를 악용한 허위 콘텐츠는 정교하고 설득력이 높아 일반 사용자가 진위를 구별하기 어렵고, 이러한 콘텐츠가 빠르게 확산되면 사회적 혼란을 초래하고 여론을 왜곡할 수 있다.
  - 예시: 선거 기간 중 AI가 생성한 가짜 뉴스나 여론 조작 콘텐츠가 유포되어 민주적 절차가 왜곡될 수 있고, 허위 재난 등의 확산은 공포와 혼란을 유발할 수 있다.
- 가짜 정보 확산 방지를 위한 윤리적 책임
  - AI 허위 콘텐츠를 공유하거나 판매하는 행위는 정보의 진실성을 해치며 윤리적 책임을 저버리는 것이다.
  - AI 시대에서 이용자는 허위 정보 확산 방지를 위해 신중하게 행동해야 할 책임이 있다.

이용자 주의 사항

- 콘텐츠 진위 확인 및 출처 검증
  - 허위 콘텐츠는 진짜처럼 보이도록 정교하게 만들어지므로 항상 진위 여부를 확인해야 하고, 콘텐츠의 출처가 신뢰할 수 있는 기관이나 출처인지 검증한다.
- 의심스러운 콘텐츠 저장 및 공유 금지
  - AI 악용 콘텐츠(딥페이크, 가짜 뉴스 등)는 불법일 가능성이 높으므로 소장하거나 공유하지 않아야 한다. 의심스러운 콘텐츠는 즉시 삭제하고, 불필요하게 유포되지 않도록 차단한다.
- 불법 콘텐츠 유포의 법적 책임 인지
  - AI 허위 콘텐츠를 매매하거나 공유하면 법적 처벌 대상이 될 수 있으므로 항상 법적 책임을 인지해야 한다.

보안 위협

- 사이버 범죄 확산: 허위 콘텐츠를 매매하거나 공유하면 악의적인 행위자들이 이를 범죄 목적으로 활용할 가능성이 높아진다.(예시: 딥페이크를 사용한 사기, 협박, 금전 요구(블랙메일) 등)
- 사회적 혼란 및 신뢰 붕괴: 허위 콘텐츠가 정치적, 사회적, 경제적 이슈에 대해 잘못된 정보를 퍼뜨리면 공공 혼란과 불신을 조장할 수 있다.(예시: 정치인의 허위 발언을 담은 딥페이크 영상이 대중에게 확산되어 선거 결과에 영향을 미침)
- 데이터 유출 및 보안 침해: AI로 생성된 허위 콘텐츠가 악성코드 또는 피싱 수단으로 활용될 경우, 이를 매매하거나 공유하면 광범위한 데이터 유출 및 보안 침해가 발생할 수 있다.(예시: 악성 첨부파일이 포함된 AI 생성 이메일을 대량으로 유포하여 기업 네트워크 침투)
- 개인정보 침해 및 명예 훼손: AI가 생성한 허위 콘텐츠(딥페이크, 가짜 메시지 등)를 통해 특정 개인이나 단체의 명예를 훼손하거나 사생활을 침해할 수 있다.(예시: 허위 음란물 딥페이크를 제작 및 유포하여 피해자에게 심각한 정신적, 사회적 피해를 초래)
- 보안 시스템 악용: AI 악용 콘텐츠가 보안 시스템을 교란하거나 우회하는 도구로 사용될 수 있다.(예시: 위조된 음성이나 얼굴 데이터를 사용해 생체 인증 시스템을 우회하고, 민감한 정보에 접근 가능)
- 허위 기술 확산: 악성 AI 기술이나 허위 콘텐츠 제작 방법이 매매 및 공유를 통해 확산되면, 더 많은 공격자가 이를 활용하여 보안 위협이 증가한다.(예시: AI 기반 딥페이크 제작 소프트웨어를 공유하여 누구나 쉽게 허위 콘텐츠를 생성 가능)