AUTOCRYPT

Declaration on Cybersecurity

사이버시큐리티 선언문

2026 년 10 월 5 일

서문

AI 가 스스로 해킹하기 시작했다. 점점 더 디지털화되고 있는 우리 사회는 커다란 사이버시큐리티 위협 앞에 놓이게 되었다.

AI 의 시대에 사이버시큐리티를 올바로 갖추는 방안에 세세한 답을 내놓는 것은 매우 어려운 일이다. 모든 디지털 시스템과 AI 간의 상호작용에 수반하는 사이버 위협에 대해 정확히 예측할 수 없기 때문이다. 그보다는, 사이버시큐리티에 관한 명료하면서 일관성 있는 기준을 마련하고 이를 개별 상황에 적용하며 문제를 풀어 나아가는 것이 효과적이라 할 수 있다.

무차별적인 해킹이 커다란 사회 문제로 급격하게 부각됨에 따라 위에 이야기한 기준을 시급히 정립할 필요를 느낀 아우토크립트와 펜타시큐리티의 암호학, 소프트웨어공학, Physical AI 등 관련 전문가들이 모여 심도 있는 논의를 하였다. 우리 외에도 많은 사람들이 같은 노력을 하고 있을 것이라 생각한다. 보다 포괄적이고 효과적인 논의와 협력을 위한 시발점이 되길 바라며 그 내용을 공유한다.

I. 사이버시큐리티 선언

Key Element. 사이버시큐리티는 사회의 기본 구성 요소이다.

우리는 디지털화된 시대를 살고 있다. 모든 디지털 시스템에서 사이버시큐리티는 반드시 보장되어야 한다.

Comprehensiveness. 사이버 위협에 대한 방어 대책은 빈틈없이 갖추어져야 한다.

'공격 비용이 이익보다 크면 공격받지 않는다'는 사이버시큐리티의 전제는 AI 가 공격하는 시대에는 유효하지 않다. 모든 디지털 시스템은 AI 사이버어택의 대상이며 매우 작은 취약성만 있어도 위험하다. 따라서 사이버시큐리티의 모든 조치가 마련되어 있어야 한다. 이를 위해 **Cyber Defence 3 원칙**을 아래에 제시한다.

Security by Design. 모든 디지털 시스템은 사이버시큐리티를 우선 고려하여 설계되어야 한다.

일부 중요 IT 시스템과 자동차를 포함한 일부 Physical AI 에는 설계 단계부터 사이버시큐리티가 적용되는데, 다른 모든 디지털 시스템도 그렇게 되어야 한다. 사이버시큐리티의 개념이 시스템의 핵심 기능 설계 사상에 녹아들어 있어야 한다.

II. Cyber Defence 3 원칙

Observability. 사이버시큐리티 정보의 측정

모든 디지털 시스템에는 사이버시큐리티에 관련된 모든 데이터가 측정되고 분석될 수 있는 기술적 방안이 마련되어 있어야 한다. 이 방안은 디지털 시스템의 운영체제 및 그와 상호작용하는 AI 로부터 독립적으로 구현·운용되어야 한다.

Cross-Layer Defence. 모든 Attack Surface 에 대한 방어 체계와 이들의 협응

디지털 시스템 구성 요소들 간의 접점은 사이버시큐리티 관점에서 취약하다. 이들이 연결되면 Attack Surface 가 형성된다. 디지털 시스템의 발달에 따라 접점이 증가할수록 Attack Surface 도 넓어진다. Attack Surface 는 여러 층(Layer)으로 존재하므로 적절한 사이버시큐리티 체계가 층별로 구축되고 서로 수평적·수직적으로 협응해야 한다.

Zero Damage Design. 사이버어택의 무효화

사이버어택의 최종 목적은 중요 정보 자산의 탈취, 위·변조, 시스템의 무력화 혹은 통제권 장악이다. 따라서 이러한 목적을 무효화할 수 있는 방안이 마련되어야 한다.

- 모든 정보 자산의 암호화와 안전한 키 관리
- 핵심 기능의 Resilience 와 체계적 복구 보장
- 피해가 전이되지 않는 Compartmental Design
- 핵심 기능에 대한 Fail-Safe, Minimal Risk Maneuver, 비상정지장치(Red Button) 보장

III. 안전한 AI 를 위한 사이버시큐리티 제안

AI 의 잠재적 위험성에 대한 우려가 많은 만큼 안전한 AI 를 만들기 위한 방안에 대한 논의와 연구가 활발하다. 그러나 이러한 노력에도 불구하고 AI 가 설정된 안전 기준을 회피하거나, 안전 기준이 부주의로 혹은 '의도적(나쁜 AI)'으로 결여된 경우를 생각하지 않을 수 없다. 따라서 AI 의 사이버시큐리티 리스크를 통제하는 외부적 방법에 대하여 다음과 같이 제안한다.

Trustworthy AI:

AI 는 검증된 방식으로 일해야 한다. AI 에게 '허용된 행동'과 '허용되지 않은 행동'에 대한 엄격하고 모순 없는 정의가 있어야 한다. 정의되지 않은 행동은 '허용되지 않은 행동'으로 간주한다. AI 의 접근과 행동은 객관적으로 관찰, 기록, 설명될 수 있어야 한다.

Controllable AI:

Trustworthy AI 를 보장하기 위해 AI 의 행동에 대한 외부적 통제권이 있어야 한다. AI 의 행동에 대한 최종 권한은 사람이 가지며, 모든 행동은 언제든 거부되거나 종료될 수 있어야 한다. AI 가 작동하는 물리적 환경에 대한 인간의 완전한 통제권 역시 보장되어야 한다.

Accountable AI:

AI Agent 에는 반드시 책임지는 주체가 정의되어 있어야 한다. 문제가 발생하면 책임 주체는 신속하게 고지하며 Agent 를 수정하여야 한다. 모든 AI Agent 에는 작성자, 소유자, 목적, 종료조건이 명기되어야 한다.

IV. 미래에 대하여

AI 가 보여준 놀라운 잠재력은 큰 기대와 두려움을 동시에 일으키고 있다. 그 우려 가운데 대표적인 것이 사이버시큐리티에 대한 것인데 급속히 현실화되고 확산되고 있는 중이다. 최근 들어 온갖 해킹 사건들이 보도되고 있고 매번 AI 가 관련된 정황이 언급되고 있다. AI 는 위협인가?

AI 는 인간이 축적해 온 수학적, 공학적 지식 위에서 인간에 의해 탄생한 것으로 그 제작 원리는 잘 알려져 있다. 우리가 모르는 것은 AI 가 사회의 다른 시스템들과 어떻게 상호작용을 하며 어떤 결과를 낳을 것인가이다. 그것이 무엇이든 인류에게 해가 되지 않고 도움이 되는 방향으로 나아가야 한다. 사이버시큐리티는 이 방향성을 결정하는 데 있어 가장 결정적 영향력을 갖고 있는 요소이다. 적절한 기술 개발과 관련 제도의 수립, 사회의 전면적인 수용이 필요한 시점이다.

공동 선언인

이석우
아우토크립트 의장/창업자, 소프트웨어공학, M.E.

김덕수
아우토크립트 CEO/공동창업자, 암호학, M.E.

김태균
펜타시큐리티 CEO, 해군정보통신장교

심상규
아우토크립트 CTO/공동창업자, 암호학, Ph.D.

정성균
아우토크립트 엔지니어링본부장, 정보보안, M.E.

정태준
펜타시큐리티 기획실장, 보안소프트웨어전략

이준용
펜타시큐리티 연구소장, 보안소프트웨어개발

조성우
아우토크립트 AI 모빌리티정책연구소장, 자동차 AI, 기술사/Ph.D.

부록: 용어 정의

디지털 시스템: 소프트웨어에 의해 동작하며, 데이터를 처리하거나 다른 시스템과 연결되어 기능을 수행하는 시스템

Physical AI: AI 가 물리적 환경을 인지·판단하고, 그 결과에 따라 실제 세계에서 직접 작동하는 시스템

Attack Surface: 공격자가 시스템에 침입하거나, 기능에 영향을 주거나, 데이터를 빼낼 수 있는 지점의 집합. 외부 연결 지점과 구성 요소 간의 접점을 모두 포함

Resilience: 공격이나 장애에 대비하고, 그 상황에서도 핵심 기능을 유지하며, 신속하게 복구하고 적응하는 능력

Compartmental Design: 시스템을 서로 격리된 구획으로 나누고 구획 간 연결을 통제된 경로로 제한하여, 한 구획의 침해가 다른 구획으로 전이되지 않도록 하는 설계 방식

Fail-Safe: 고장이나 침해가 발생했을 때 피해가 생기지 않도록, 시스템을 사전에 정의된 안전한 상태로 전환하는 설계 원칙

Minimal Risk Maneuver: 정상적인 작동을 계속할 수 없을 때, 시스템이 스스로 위험이 최소화된 상태로 전환하는 절차

Red Button: 사람이 한 번의 조작으로 AI 또는 시스템의 작동을 중단시켜 안전한 상태로 멈추게 하는 비상정지 수단. 물리적 장치와 소프트웨어 절차를 모두 포함

AI Agent: 주어진 목표를 달성하기 위해 작업 절차를 스스로 판단하고, 도구나 다른 시스템을 사용하여 행동하는 AI

AUTOCRYPT

아우토크립트 주식회사 · 서울특별시 영등포구 여의공원로 115, 07241 · www.autocrypt.io

© 2026 AUTOCRYPT Co., Ltd. All rights reserved.