시민사회를 위한 생성형 AI 가이드

디지털정의네트워크 사회적협동조합 빠띠 정보인권연구소

시민사회를 위한 생성형 AI 가이드

발행일 2025년 12월
글쓴이 오병일·고아침·장여경
펴낸곳 디지털정의네트워크 · 사회적협동조합 빠띠 · 정보인권연구소
후원 진보통신연합(APC) www.apc.org
편집 다디잔
주소 03745 서울시 서대문구 독립문로8길 23 (천연동, 3층)
전화 02-774-4551
팩스 02-701-7104
홈페이지 guide.digitaljustice.kr/generative-AI

시민사회를 위한
생성형 AI 가이드

디지털정의네트워크
사회적협동조합 빠띠
정보인권연구소

들어가며

“단체 활동가들이 챗지피티로 성명서를 작성하는데,
어떻게 해야할지 고민이에요”

아마도 많은 단체에서 이러한 고민을 하고 있을 것입니다. 챗지피티, 제미나이 같은 챗봇뿐 아니라, 이미지·음악·영상을 생성하는 다양한 생성형 AI 서비스들이 등장하면서, 점점 더 많은 시민들이 업무와 개인적인 목적으로 이를 사용하고 있습니다. 시민사회 활동가 역시 예외는 아닙니다.
그런데 아직까지는 업무 목적이더라도 개인의 판단에 따라 활용하고 있을 뿐, 조직 차원의 생성형 AI 정책을 갖춘 단체는 거의 없는 상황입니다.

시민사회단체가 생성형 AI를 사용할 때 주의해야 할 점들이 많습니다. 예를 들어, 생성형 AI의 할루시네이션으로 인한 사실이 아닌 내용이 단체의 공식 문서에 포함될 경우 단체의 신뢰가 심각하게 훼손될 수 있습니다. 신뢰할 수 없는 상용 서비스에 개인정보나 기밀정보를 업로드할 경우 보안에 문제가 생길 수 있습니다. 생성형 AI의 결과물이 단체의 가치와 맞지 않는 편향을 포함할 수도 있습니다. 생성형

AI로 성명서를 작성하는 과정에서 활동가의 역량 강화와 단체 내 숙고라는 측면은 배제될 수 있습니다. 조직 차원의 정책없이 활동가 개인의 선택에 따라 AI 도구를 사용할 경우 조직이 통제할 수 없는 문제가 발생할 가능성이 큽니다.

그러나 시민사회단체가 생성형 AI 서비스를 활용하는 것이 적절한지, 활용할 경우 어떠한 원칙과 정책 하에 활용하는 것이 바람직한지, 인권의 관점에서 참고할 수 있는 지침은 많지 않습니다. 시민사회단체 활동가들이 어떤 업무를 위해 어떤 AI 도구를 사용하고 있는지에 대한 현황도 파악되어 있지 않은 상황입니다. 이 가이드는 시민사회단체와 활동가들이 생성형 AI 정책을 수립하고, 필요할 경우 올바로 활용할 수 있도록 돕기 위한 문제의식에서 출발했습니다.

가이드를 만들기 위해, 실제 어떤 업무에 어떤 AI 도구를 사용하고 있는지, 생성형 AI가 얼마나 유용한지, 사용하면서 경험하는 문제는 무엇인지에 대해 설문조사를 진행했습니다. 국내 활동가들 뿐만 아니라 APC 네트워크를 통해 전 세계 활동가들의 의견도 수렴하였습니다. 응답 규모가 크지 않아 통계적인 의미는 제한적이지만 활동가들이 느끼는 실제 고민과 공통된 문제의식을 확인할 수 있었습니다. 생성형 AI를 거의 사용하지 않는 분들도 응답을 해서 자신의 생각을 공유해 주었습니다.

또한 시민사회, 노동조합 활동가들과 함께 생성형 AI를 주제로 한 워크숍을 열어 설문조사 결과와 정책 모델 초안을 공유하고 서로의 경험과 생각을 나누었습니다. 이 과정에서 합의된 결론을 도출하는 것보다 서로의 느낌과 고민을 솔직하게 나누는 과정 자체가 중요하다는 점을 다시 확인하였습니다. 이 가이드가 제시하는 정책 모델은 하나의 출발점일 뿐이며 각 단체의 현실과 활동가들의 목소리를 반영하여 스스로 정책을 만들어가는 과정이 무엇보다 중요합니다.

생성형 AI를 흥미롭게 사용하는 활동가들도 있지만 여전히 생성형 AI 자체를 불편하게 느끼는 분들도 많습니다. 이 가이드는 생성형 AI의 활용을 권장하려는 목적이 아니라는 점을 분명히 밝힙니다. 주요 생성형 AI 모델의 개발과 서비스의 제공이 빅테크 기업에 의해 독점적으로 이루어지고 있다는 것도 문제입니다. 이 가이드는 현재 주로 사용되는 상용 생성형 AI 서비스에 초점을 두고 있지만, 이러한 구조적인 한계를 극복해야 한다는 문제의식에 깊게 공감합니다.

여러 한계에도 불구하고 이 가이드가 현재 생성형 AI 관련 정책을 고민하는 단체와 활동가들에게 조금이나마 도움이 되기를 바랍니다.

차례

1장. 생성형 AI 관련 주요 개념

이 장에서는 생성형 AI를 둘러싼 개념 몇 가지를 소개합니다.
처음부터 읽어도 되고, 궁금한 키워드가 있을 때
하나씩 살펴보는 용도로 이용해도 좋습니다.

인공지능(AI, Artificial Intelligence)

'인공지능'은 다양한 의미로 쓰이는, 꽤 느슨한 개념입니다. 컴퓨터 과학의 세부 분과로서 인공지능은 학습/추론/지각 등 인간의 지적 능력을 인공적으로 구현하는 것을 목적으로 합니다. 그런 목적으로 구현된 시스템이나, 구현을 위한 방법론 등을 가리키기도 합니다.

2026년 1월 시행되는 인공지능기본법에서는 인공지능을 다음과 같이 정의합니다.

- 인공지능 : 학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것
- 인공지능시스템 : 다양한 수준의 자율성과 적응성을 가지고 주어진 목표를 위하여 실제 및 가상환경에 영향을 미치는 예측, 추천, 결정 등의 결과물을 추론하는 인공지능 기반 시스템

우리가 일상 대화에서 이야기하는 '인공지능'은 보통 개별적으로 구현된 시스템(예: 챗지피티)이나, AI 기술 분야 전반을 지칭합니다. 최근 통용되는 '인공지능'은 생성형 AI 기술을 가리키는 경우가 많지만, 추천 시스템이나 채용 알고리즘 등 비-생성형 기계학습 기술도 인공지능에 포함됩니다. 그런가 하면, 지도에서 최적의 이동경로를 찾는 일은 한때 인공지능 분야의 중요한 문제였지만 오늘날

지도 앱의 '길찾기' 기능을 보고 '인공지능'이라고 표현하는 경우는 많지 않습니다. 이처럼 무엇을 '인공지능'이라고 부르는지는 시대와 맥락에 따라 달라지기에 용어 사용시 그 지칭 대상을 의식적으로 명확히 할 필요가 있습니다.

- **기계학습 (머신러닝 Machine Learning)**

생성형 인공지능은 기계학습 기법으로 구현합니다. 그렇다면 기계학습은 무엇일까요? 데이터를 바탕으로 통계 알고리즘('모델')을 '학습'하고, 학습한 적 없는 데이터를 처리할 수 있는 능력을 확보(일반화)하여 명시적 지침 없이도 기능을 수행할 수 있게끔 하는 일련의 기법으로 정의할 수 있습니다.

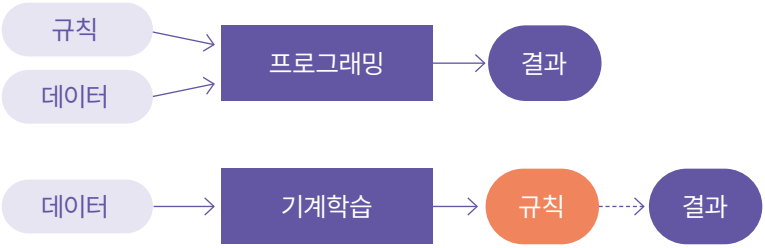

그림 1. 일반 프로그래밍과 기계학습의 차이

기계학습은 컴퓨터가 데이터를 바탕으로 패턴을 '학습'해 예측이나 결정을 내리는 기술입니다. 예를 들어, 음악 추천

서비스가 사용자의 취향을 알아내거나, 스팸 메일 필터가 정상 메일을 구분하는 것이 기계학습의 결과물입니다. 명시적인 프로그래밍 없이 데이터 자체에서 규칙을 찾아내는 '데이터 기반 자동화 기술'이라고도 할 수 있습니다.

- **인공신경망 (Artificial Neural Network)**

기계학습 알고리즘의 한 유형입니다. 인간의 뇌가 뉴런을 연결해 정보를 처리하는 방식에서 영감을 받았지만, 실제 생물학적 구조와는 차이가 있습니다. 수많은 단순한 처리 단위('뉴런', 일종의 간단한 함수)가 계층적으로 연결되어 커다란 하나의 함수(신경망)를 구성합니다. 뉴런 간의 연결은 각각 가중치를 가지며, 이 가중치를 조정해 패턴을 인식하거나 예측하는 능력을 개선하는 과정을 '학습'이라고 부릅니다.

- **딥러닝 (Deep Learning)**

인공신경망을 여러 겹으로 쌓아 복잡한 패턴을 처리하는 기계학습 기법입니다. 딥러닝(심층학습)이란 이름은 인공신경망이 여러 층(레이어)으로 구성되어 있는 구조를 지칭합니다. 이같은 다층 구조 설계와 대규모 데이터 학습을 통해, 이미지 인식이나 긴 텍스트 속 단어 간 관계를 파악하는 등 기존에는 인공지능으로 처리하기 어려웠던 과업에 대한 성능이 현저히 개선되었습니다. 생성형 인공지능은 대부분 딥러닝을 기반으로 작동합니다.

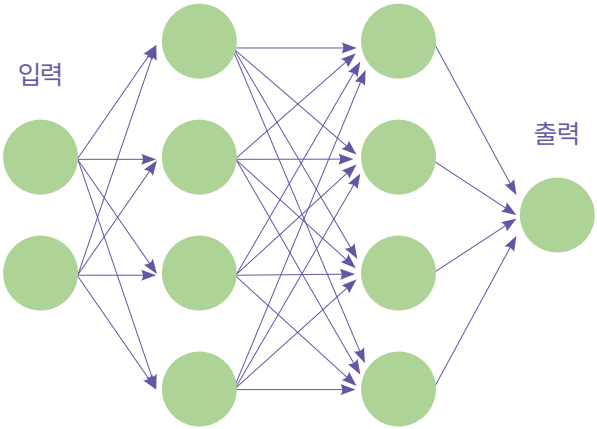

그림 2. 단순화한 딥러닝 모델 구조 예시

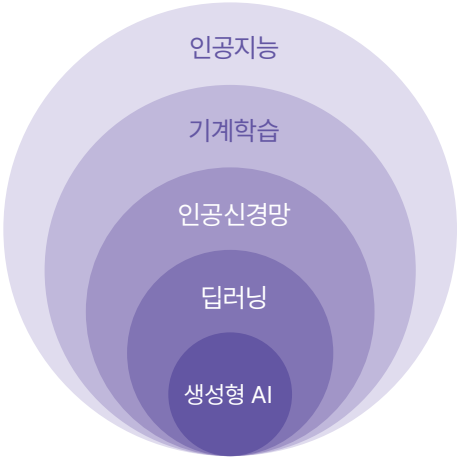

그림 3. 인공지능의 여러 유형 간의 관계

- **생성형 인공지능 (Generative AI)**

  생성형 인공지능은 이용자의 입력('프롬프트')을 바탕으로 텍스트, 이미지, 음성, 영상 등 콘텐츠를 만드는 기술입니다. 생성형 인공지능은 대량의 데이터에 기계학습 기법을 적용하여 만든 '모델'을 엔진 삼아 작동하며, 널리 쓰이는 상용 생성형 인공지능 도구에서는 보통 채팅 인터페이스를 통해 이러한 모델과 상호작용하게 됩니다.

- **자연어처리 (NLP, Natural Language Processing)**

  인간의 언어(텍스트 또는 음성)를 컴퓨터를 활용해 이해·해석·생성할 수 있도록 하는 분야를 자연어처리라 부릅니다. 구문 분석, 기계번역, 고유명사 등 개체명 인식, 음성 인식 등 다양한 자연어처리 과업 유형이 있으며, 챗봇 등의 생성형 인공지능이 수행하는 기능 즉 대량의 텍스트 데이터를 학습해 문맥에 맞는 자연스러운 글을 생성하는 일 또한 자연어처리의 한 예시로 이해할 수 있습니다.

- **거대언어모델 (LLM, Large Language Model)**

  거대언어모델은 방대한 텍스트 데이터를 학습해 언어를 해석하거나 생성하는 인공지능 모델입니다. 생성형 AI의 핵심 기술로, 일반적으로 수십억 개 이상의

매개변수(연결 가중치)를 지닌 인공신경망을 활용해 문장 내 단어 간 복잡한 패턴을 학습합니다. 예를 들어, “오늘 날씨가”라는 입력에 “맑습니다”라는 단어를 예측하듯, 문맥을 분석해 자연스러운 문장을 완성하거나 질문에 답변하는 기능을 수행합니다. 챗지피티(ChatGPT) 같은 생성형 AI 서비스의 기반이 되는 기술입니다.

- **파운데이션 모델**

파운데이션 모델(기반 모델)은 방대한 데이터로 사전 학습되어 다양한 작업에 활용할 수 있는 대규모 인공지능 모델을 부르는 용어입니다. 생성형 인공지능 분야에서 텍스트 생성, 이미지 제작, 음성 합성 등 특정 과업에 맞춰 미세 조정(fine-tuning)되기 전 ‘기반’이 되는 모델을 의미합니다. 예를 들어, GPT-4나 Stable Diffusion은 각각 언어·이미지 생성을 위한 대표적인 파운데이션 모델로, 챗지피티 등 생성형 AI 기반 서비스는 이들 모델을 응용하여 작동합니다.

- **멀티모달**

텍스트, 이미지, 음성, 영상 등 서로 다른 형태의 데이터를 동시에 처리하거나 생성하는 기능을 의미합니다. 미드저니처럼 텍스트 설명을 기반으로 이미지를 생성하는 AI나 반대로 이미지에 담긴 내용을 텍스트로 묘사하는 기능, 음성 명령을

이해해 영상 콘텐츠를 추천하는 시스템 등이 멀티모달 접근 방식을 취하는 사례입니다. 모달리티(modality)는 기호학에서 글/이미지/음악 등의 양태를 가리키는 용어인데, AI 맥락에서는 '데이터 형식'과 비슷한 뜻으로 이해할 수 있습니다. 멀티모달 모델은 다양한 데이터 유형 간 관계를 학습함으로써 단일 데이터 형식만 다루는 모델으로는 수행할 수 없는 종류의 작업을 수행합니다.

- **강화 학습**

강화 학습은 인공지능 시스템이 환경과의 상호작용을 통해 보상을 최대화하는 결정을 내리는 기계학습 방법입니다. '에이전트'라고 부르는 시스템이 주어진 상태에서 특정 행동을 선택하고, 그 결과로 보상(성공 시 긍정적, 실패 시 부정적)을 받으며, 이를 반복함으로써 최적의 전략을 학습합니다. 게임 AI가 점수를 높이기 위해 전략을 개선하거나 로봇이 물체 집기 동작을 연습하여 성공 확률을 높이는 과정이 이에 해당하며, 이세돌 기사와의 승부로 화제가 된 알파고도 강화 학습 시스템입니다.

생성형 AI에서는 강화 학습이 생성된 출력의 품질을 개선하는 데 활용될 수 있습니다. 예컨대 사용자의 피드백을 보상 신호로 삼아 챗봇이 더 자연스럽운 답변을 생성하도록 혹은 혐오 발언을 내놓지 않도록 훈련하거나, 이미지 생성

모델이 특정 스타일 기준에 부합하는 결과를 만들 수 있도록 조정하는 데 적용됩니다. 학습 데이터의 패턴을 모방하는 것에 더해 외부 평가 기준에 맞추어 생성 능력을 최적화하는 데 활용하는 것으로, 이처럼 생성형 AI 시스템을 미세 조정하는 작업을 RLHF(Reinforcement Learning from Human Feedback, 인간 피드백 기반 강화 학습)라고 부릅니다.

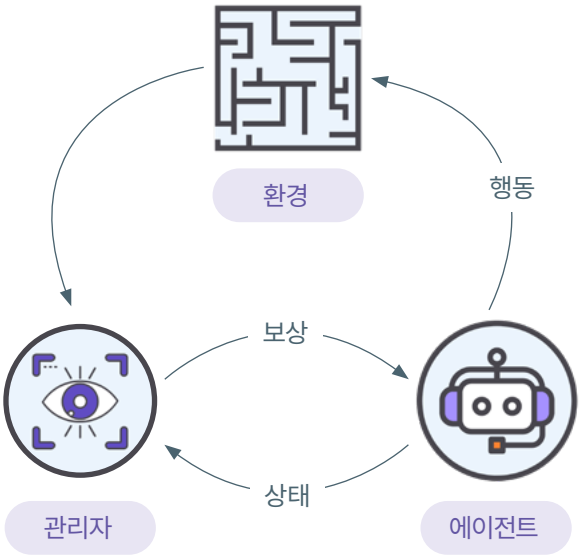

그림 4. 강화 학습 시스템의 구성 요소
출처 commons.wikimedia.org/wiki/File:Reinforcement_learning_diagram.svg

트랜스포머

트랜스포머(Transformer)는 인공신경망 설계·구현 방식의 한 가지입니다. 트랜스포머 이전에 주로 쓰이던 LLM 구현 방식에서는 학습 데이터를 순차적으로 입력했습니다. 예를 들어 "대한민국의 모든 권력은 국민으로부터 나온다"라는 문장이 있다면 "대한민국 → 의 → 모든 → ..." 순으로 입력하여 앞뒤 단어 간의 관계를 파악합니다. 이같은 기존 방식의 한 가지 문제는 장기 의존성(멀리 떨어진 토큰 간의 관계)을 포착하기 어렵다는 것인데요. 단순화하자면 앞 문장에서 "대한민국"과 "의" 사이의 연결은 명확하지만, 상대적으로 멀리 떨어진 "대한민국"과 "국민" 사이의 연결은 불분명하다는 식입니다.

트랜스포머에서는 학습 데이터를 한번에 병렬로 입력하여 장기 의존성 문제에 대응합니다. 앞 단어와의 관계만을 고려하는 것이 아니라 문장 전체에서 어느 단어와 관계가 높은지 수치화하는 것으로, 이런 기법을 자기 주의(self-attention) 메커니즘, 간단히는 어텐션이라고 부릅니다. 이러한 트랜스포머는 현재 텍스트 기반 생성형 인공지능의 핵심 기술 중 하나로, 오픈AI가 개발한 거대언어모델 시리즈인 GPT(Generative Pre-trained Transformer) 등 많은 모델이 트랜스포머 구조를 기반으로 작동합니다.

에이전트(agent)

컴퓨터과학 분야에서 '에이전트'는 각종 자동화 프로그램 및 시스템을 지칭할 수 있는 표현입니다. 생성형 AI에서 말하는 에이전트는 콘텐츠 생성과 환경 상호작용을 결합해 특정 목표를 수행하는 시스템을 의미합니다. 즉 질문에 답변을 생성하기만 하는 것이 아니라 다른 프로그램, 데이터베이스 등과 연동되어 추가적인 자동화 과정을 수행합니다. 예를 들어, 여행 계획을 돕는 에이전트는 (통상적인 채팅 기반 LLM처럼) 일정을 기획하는 것에 더해 항공권 예약 API를 호출하거나, 현지 정보를 검색해 맞춤형 추천을 제공할 수 있습니다.

- **환각(hallucination)**

생성형 AI의 '환각'은 허구이거나 오해의 소지가 있는 내용, 이용자의 의도와 무관한 내용 등을 생성하여 사실인 것처럼 제시하는 현상을 말합니다. 예컨대 텍스트 생성 시 사실과 무관한 주장을 하거나, 이미지 생성 시 입력 설명에 없는 사물을 추가하는 경우 등입니다. 생성형 AI가 답변을 생성하는 방식은 데이터 패턴 기반의 통계적 예측으로, 사실 여부에 대한 판단은 생성형 AI의 작동 방식과 무관하기에 발생하는 현상입니다. 생성형 AI의 답변 생성은 '사실' 여부와 무관하다는 의미에서 생성형 AI의 모든 출력은 일종의 '환각'이라고 볼 수도 있습니다만, 통상적인 맥락에서 AI 환각은 허위 정보 등 사실에 부합하지 않는 결과물을 가리키는 표현입니다. 한편 '환각'이라는 표현은

AI 시스템이 마치 감각적 경험을 하는 것처럼 그 작동을 과도하게 의인화하기 때문에 부적절하며, '거짓 정보'나 '헛소리' 등이 더 어울린다고 보는 시각도 있습니다.

- **RAG**

RAG(Retrieval-Augmented Generation, 검색 증강 생성, 래그)는 생성형 AI의 특징이자 단점인 환각(hallucination, 사실이 아닌 내용 생성)을 보완하여 모델의 정확성을 높이기 위한 기법입니다. RAG는 먼저 사용자의 질문에 관련된 정보를 외부 데이터베이스나 시스템이 보유한 문서에서 검색하고, 이후 검색된 정보를 바탕으로 답변을 생성하는 식으로 작동합니다. 이를 통해 특히 최신 정보나 특정 분야의 전문 지식을 반영한 답변의 정확성을 높일 수 있습니다. 단 이때도 답변을 생성하는 단계에서 오류가 발생할 가능성이 남아 있기 때문에 출처를 검증할 필요가 있습니다. RAG의 예로는 구글 등 검색엔진에 도입된 AI 요약 답변이 있습니다.

- **매개변수(parameter)**

AI 모델의 매개변수는 모델 작동에 영향을 주는 각종 내부 수치를 말합니다. 신경망 기반 모델에서는 뉴런 간 연결의 가중치(강도) 및 편향을 매개변수로 사용하며, 데이터 학습 과정에서 각 매개변수의 값을 조금씩 조정함으로써 성능을

개선합니다. 수많은 다이얼이 달려 있는 거대한 제어반에서 각 다이얼의 위치를 조금씩 돌리는 장면을 상상하면 비슷할 것 같습니다. 생성형 AI 모델은 수십억에서 수조 개의 매개변수를 통해 단어 간 복잡한 관계를 학습하고, 이를 바탕으로 문장을 생성하거나 질문에 답변합니다. 매개변수의 규모가 클수록 모델은 더 정교한 패턴을 포착하고 성능이 향상될 수 있지만, 학습 데이터와 컴퓨팅 자원에 대한 의존도도 함께 증가합니다. 신경망의 가중치과 편향 외에도 데이터 학습과 예측에 영향을 미치는 '초매개변수(hyper-parameter)'들이 있습니다. 가중치와 편향 등의 매개변수는 학습 과정에서 자동으로 계산되는 값인 반면, 초매개변수는 모델 제작자/사용자가 직접 지정하는 값입니다.

- **가중치와 편향**

가중치와 편향은 각종 기계학습 모델의 핵심 구성 요소입니다. 어떤 AI 모델이 '수십억 개의 매개변수를 가지고 있다'고 할 때 매개변수가 가중치와 편향을 가리킵니다.

입력 x와 출력 y의 관계를 표현하는 함수 y = wx + b가 있다고 할 때, x에 곱하는 값 w는 가중치, x에 더하는 값 b는 편향에 해당합니다. 가중치는 모델이 입력 데이터를 (정확히는 입력 데이터의 특정 패턴이나 특징을) 얼마나 중요하게 다루는지 결정합니다. 편향은 입력 데이터와 무관하게

모델의 작용을 일정 방향으로 끌고 가는, 일종의 '기준점' 역할을 합니다. AI 모델의 '학습'은 주어진 데이터 패턴에 부합하도록 이 가중치와 편향을 조정하는 과정입니다.

인공신경망의 예를 들어봅시다. 인공신경망은 '뉴런'을 여러 층에 걸쳐 연결해놓은 하나의 거대한 계산식입니다. 보통 특정 층에 있는 모든 뉴런은 다음 층에 있는 모든 뉴런과 연결되어 있습니다. 즉 각 뉴런은 이전 층에 속한 모든 뉴런의 출력의 합을 입력으로 받습니다. 뉴런 사이의 연결마다 다른 가중치가 적용됩니다. 각 뉴런은 하나 이상의 숫자 입력(이전 층에 속한 뉴런의 출력의 합)을 받아 출력을 내놓는 함수로, 뉴런마다 편향 값을 가집니다. 이 편향을 합산한 뒤 나온 값에 따라 다음 층으로 전달되는 정보가 결정됩니다.

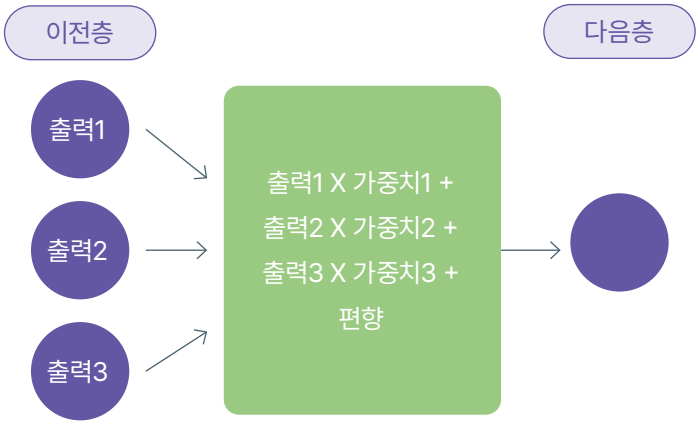

그림 5. 개별 뉴런의 작동 방식

한편 생성형 AI에서 '편향'은 모델의 학습 데이터나 설계 과정에 기인하여 특정 집단이나 관점을 과도하게 반영하거나 배제하는 현상을 의미하기도 합니다. 예를 들어 텍스트를 생성할 때 특정 직업군을 특정 성별이나 인종과 연관시킨다거나, 역사적 서술을 특정 집단의 관점으로만 한다거나 하는 등의 사례가 있을 수 있습니다. AI 시스템의 편향은 학습 데이터에 내재된 사회적 고정관념이나 데이터 수집의 불균형을 포함해 시스템 구축의 여러 단계에서 비롯될 수 있으며, 그 활용 과정에서 이러한 편향을 재생산할 수도 있습니다. 이에 대응하기 위해 학습 데이터의 다양성 확보, 알고리즘 개선, 지속적인 모니터링과 같은 기술적·윤리적 접근을 시도하기도 합니다.

다만 개별 뉴런의 편향이 학습 과정에서 결정된 하나의 상수이듯, 특정 AI 시스템의 편향 또한 그 구축 과정에서 비롯된 일종의 위치성이라고 볼 수 있기에 편향을 완전히 '제거'한 AI 시스템은 불가능한 목표라고 할 수 있습니다. 다만 소수자 배제 등의 편향이 일어나지 않도록 기술 시스템을 설계해야 한다고는 할 수 있겠습니다.

- **온도**

온도(Temperature)는 생성형 AI 모델의 출력 다양성(무작위성)을 조절하는 초매개변수입니다. 텍스트 생성 모델의 경우 다음 단어를 예측할 때, 온도를 높게

설정하면 덜 유력한 단어도 선택할 가능성을 높여서 상대적으로 자유롭게 결과물을 생성합니다. 반대로 온도를 낮게 설정하면 가장 유력한 단어 위주로 선택하여 상대적으로 예측 가능한 출력을 생성합니다. 즉 온도값은 생성형 AI 모델의 출력 스타일을 조정하며, 예를 들어 가사 쓰기나 브레인스토밍에는 높은 온도가, 서류 작성에는 낮은 온도가 어울릴 수 있습니다. 달리 말하자면 온도를 높게 설정하면 환각의 가능성이 올라가고, 온도를 낮게 설정하면 학습 데이터 패턴을 반복할 가능성이 올라갑니다.

- **프롬프트**

생성형 AI 모델에 입력하는 데이터 내지 지시문을 프롬프트라고 합니다. 텍스트 생성 모델에서는 “문서를 요약해줘” 같은 문장, 이미지 생성 모델에서는 “푸른 바다를 배경으로 한 황금 태양 일러스트” 같은 설명이 프롬프트에 해당합니다. 멀티모달 시스템의 경우 텍스트와 이미지 등을 조합해 복합적인 프롬프트로 활용하기도 합니다.

프롬프트 엔지니어링은 원하는 출력을 얻기 위해 입력(프롬프트)을 설계하는 기법입니다. 예를 들어, “설명을 간단히 해줘”와 “초등학생도 이해할 수 있도록 비유와 예시를 포함해 설명해줘”라는 프롬프트는 동일한 질문에 대해 전혀 다른 답을 유도합니다. 지시문·역할

설정·예시·출력 형식 지정 등을 통해 생성형 AI를
효과적으로 활용하기 위한 일련의 전략에 해당합니다.

- **토큰**

자연어처리에서 '토큰'은 컴퓨터가 텍스트 데이터를 처리하는
최소 단위를 의미합니다. 이때 하나의 텍스트 조각은 단어나
음절·글자, 한국어의 경우에는 형태소 등이 될 수 있습니다.
예를 들어, "사과"라는 단어는 하나의 토큰으로 처리될 수 있고,
"unhappy" 같은 단어는 "un"과 "happy" 두 토큰으로 분리될
수 있습니다. 토큰화(tokenization) 방식은 모델마다 다르며,
같은 문장도 모델에 따라 토큰 수가 다를 수 있습니다.
보통 생성형 AI 서비스는 이용량을 토큰 단위로 계산하기에,
이용자 입장에서는 비용과 직결되기도 합니다.

- **임베딩**

임베딩은 텍스트, 이미지, 음성 등의 데이터를 컴퓨터가
처리할 수 있는 숫자 배열(벡터)로 변환하는 기법 혹은
그렇게 변환해놓은 숫자 배열을 말합니다. AI 모델은
단어, 문장, 이미지 등을 인식하거나 생성할 때 임베딩으로
변환하여 처리합니다. 임베딩은 데이터의 복잡한 패턴을
단순화하고 관계를 명확히 하는 일종의 번역기 역할을 하며,
특히 멀티모달 모델에서 텍스트-이미지 간 변환 등 다양한

작업의 기반이 됩니다. 임베딩으로 표현한 데이터는 그 의미나 특징을 수치적으로 나타내며, 이를 활용해 예컨대 '강아지'와 '고양이'라는 단어 사이의 거리가 '강아지'와 '선풍기' 사이의 거리보다 가깝고 따라서 '강아지'와 '고양이'는 유사한 개념이라는 식의 계산을 할 수 있습니다.

- **창발성**

창발성(Emergence)은 복잡한 시스템이 단순한 구성 요소의 상호작용을 통해 설계 단계에서 예상치 못한 새로운 특성이나 능력을 나타내는 현상을 의미합니다. 생성형 AI에서는 거대언어모델이 방대한 데이터와 수많은 매개변수를 학습한 뒤, 개발자가 명시적으로 프로그래밍한 적 없는 능력이 나타나는 (것처럼 보이는) 현상을 지칭합니다. 예를 들어 특정 질문에 대한 논리적 답변 생성이나 다양한 스타일로의 텍스트 변환 능력 등이 창발성의 결과로 이야기됩니다. 긍정적인 특징뿐만 아니라 허위정보, 해킹 등의 부정적 방향으로도 나타납니다. 생성형 AI의 이같은 창발적 특성이 정말 '새로운' 능력인지는 아직 논쟁의 대상입니다만, 우리가 생성형 AI 기술을 이해하고 그 작용을 예측하는 데 아직 공백이 있다고 볼 수 있습니다.

- **의인화**

생성형 AI의 과도한 의인화는 기술의 본질과 한계를 흐리며

여러 위험을 야기하기에, 생성형 AI가 의도성·감정·의식이 없는 통계적 패턴 생성 도구임을 명확히 할 필요가 있습니다. 의인화는 우선 기술의 현재 수준에 부합하지 않는 과대평가와 오남용으로 이어질 수 있습니다. 감정적 유대감을 느끼고 인간처럼 발화하도록 설계된 챗봇은 사용자가 기술적 한계를 넘어 인간적 판단을 기대하게 만들고, 의료·법조 등 고위험 분야에서조차 AI에 대한 맹목적 신뢰를 유발할 수 있습니다. 또한 AI를 인간과 동등한 존재로 오인할 경우, AI로 인한 인류의 멸망이나 AI의 인격 등 비현실적 담론이 확대되며, 실제 인간 노동자의 가치나 권리, 그리고 AI 시스템을 개발하고 운영하는 기업의 책임에 관한 논의를 방해할 여지가 있습니다.

- **AGI(인공일반지능)**

AGI는 텍스트 생성, 이미지 생성 등 특정 작업에 특화된 현재 인공지능과 달리, 인간이 수행하는 다양한 지적 활동(문제의 이해·해결·추론·창의적 사고 등)을 폭넓게 수행할 수 있는 가상의 시스템을 의미합니다. 만약 AGI가 달성된다면 다양한 분야의 지식을 통합적으로 이해하고, 새로운 상황에 유연하게 대응하며, 스스로 복잡한 문제를 해결할 수 있을 것으로 기대됩니다. 그러나 현재의 기술 상황에서는 실현되지 않았고 전망도 불투명한 개념이며, 기술 용어보다는 마케팅 용어로 쓰이는 경향이 있습니다.

데이터센터

생성형 AI 모델을 제작하고 서비스를 구동하기 위해서는 고성능 GPU 서버를 대량으로 작동시켜야 합니다. 이를 위해 만들어진 데이터센터는 방대한 양의 데이터와 연산을 처리하기 위해 서버를 적게는 수천 대, 많게는 수십만 대 이상 모아둔 건물입니다. 데이터센터는 생성형 AI 산업에 필수적인 인프라인 동시에 에너지, 냉각수, 광물자원 등 각종 환경 비용과 직결되는 요인이기도 합니다.

2장. 생성형 AI 활용 관련 주요 이슈

생성형 AI의 개발과 활용은 다양한 사회적 층위에서 새로운 문제를 발생시키거나 기존 문제를 재생산합니다. 이 장에서는 생성형 AI와 노동, 환경, 보안 문제 등 기술과 사회의 관계에 있어 함께 고민해보면 좋을 몇 가지 논의점을 소개합니다.

생성형 AI 구축의 저임금 노동 착취 구조

생성형 AI 도구는 단순히 '많은 데이터를 연산하여 산출(학습)한 모델'로 완성되지 않습니다. 많은 경우 데이터를 유용한 형태로 정리하는 데이터 라벨링 작업이 필요하며, 완성된 모델은 학습 데이터 속 편향이나 오류, 혐오표현 등을 반영하고 있기에 실제 서비스를 제공하려면 부적합한 결과물 산출을 최소화하는 미세조정(fine-tuning) 작업이 필요합니다. 미세조정 역시 일종의 데이터 라벨링으로, 많은 사람이 투입되는 대규모 미세노동(microwork) 형태를 취합니다.

데이터 라벨링 노동은 미세노동 특유의 불안정성과 혐오표현 등을 마주해야 하는 정신적 부담, 많은 경우 노동력이 싼 남반구 지역으로 작업을 외주화하는 구조 등을 특징으로 합니다. 이같은 데이터 라벨링 노동 과정은 복잡한 하청 구조를 거치고 기업비밀을 이유로 은폐되기 때문에 그 현황을 정확히 파악하기도 어렵습니다. 즉 생성형 AI 도구의 생산 과정에는 복합적인 노동 착취 기제가 깔려 있다는 점에서 생성형 AI 활용은 그 생산 과정을 둘러싼 윤리적 문제를 던집니다.

자동화, 일자리 대체, 그리고 생산성

생성형 AI는 흔히 업무 자동화 및 재편을 통해 생산성을 향상시켜 주는 기술로 인식됩니다. 이는 일자리 대체 및 기존 노동자의 역할 축소로 이어질 위험이 있습니다. 사회

전반적으로도 그렇지만 개별 조직 차원에서도 생성형 AI와 같은 자동화 기술 도입은 기존 노동 방식에 변화를 초래하며, 인력 축소 등 긴축 경영과 맞물릴 여지를 만들게 됩니다. 기술적 전환 과정에서 노동권을 취약하게 하거나 향상된 생산성의 결실을 일부가 독점하지 않으려면 어떻게 해야 할까요? 재교육 및 공정한 분배 방안 등을 우선 고민해볼 필요가 있겠습니다. 또한 조직에서 생성형 AI를 도입하는 과정에서 각 이해관계자가 의사결정에 관여할 수 있는 참여적 거버넌스가 요청됩니다.

조직 내의 생성형 AI 도입은 조직 밖에도 영향을 미칠 수 있습니다. 본 가이드 준비 과정에서 진행한 워크숍 참여자는 소속 단체에서 생성형 AI를 활용해 집회용 음악을 제작한 사례를 소개했습니다. AI 생성 음악을 이용할 경우 금전적·시간적 비용이 줄어드는 효과를 기대할 수 있지만, 그러지 않았더라면 문화예술 노동자에게 의뢰했을 일감을 대체한 것으로 해석할 수도 있을텐데요. 예전이라면 외부에 맡겼을 포스터 디자인·일러스트 등을 생성형 AI를 활용해 내부 담당자가 구현하는 경우도 흔해졌습니다. 조직 차원에서의 생산성 도모가 노동 생태계 및 주변 네트워크에 부정적 영향을 미치는 긴장관계를 볼 수 있습니다.

한편 생성형 AI가 생산성에 꼭 기여하는지 또한 비판적으로 검토할 필요가 있습니다. 반복적인 작업이나 아이디어 탐색 과정에서 속도를 높이는 도구가 될 수 있지만, 그러기 위해서는

AI로 생성한 결과의 정확성 검증, 편향 또는 오류 수정, 도구 활용 역량 습득 등에 추가 시간과 자원이 투입되어야 합니다. 기술 도입이 구성원의 역량 계발과 상충하지 않도록 업무 구조를 재편하는 노력 또한 필요할 것입니다. 바꿔 말하면 조직의 디지털 전환 과정의 일환으로 생성형 AI를 이해할 필요가 있습니다. 생성형 AI에 투자한 기업 중 95%가 그로 인한 조직 차원의 순익을 내지 못하고 있다는 설문 사례[1]에서 드러나듯 이 전환은 간단치 않습니다.

학습 데이터 저작권과 창작 노동

생성형 AI 모델을 제작하려면 막대한 양의 데이터를 확보해야 합니다. 웹상에 게시된 각종 글, 이미지, 코드 등뿐만 아니라 단행본 등 출판물을 직접 활용하기도 합니다. 이 과정에서 AI 기업들이 저작권자의 명시적 동의를 구하지도 보상을 제공하지도 않는 사례가 빈번합니다. 이러한 행태가 '공정 이용'이라는 산업계의 관점과 창작자 권리 침해라는 관점이 대립하는 가운데 TDM(Text and Data Mining)이라고도 표현하는, 저작물의 AI 학습 데이터 활용의 허용 및 규제에 관한 입법적 논의도 진행 중입니다. 법제도 수립과 별개로 학습 데이터 관련 저작권 소송도 다수 진행 중이며, 이들 소송의 향방 역시 중요한 참고점이 될 수 있습니다.

---

1 Aditya Challapally, Chris Pease, Ramesh Raskar, Pradyumna Chari. *The GenAI Divide: State of AI in Business 2025*. MIT NANDA.

생성형 AI 이용자 입장에서는 기존 저작물과 같거나 유사한 산출물을 만들어내어 타인의 저작권을 침해할 위험이 존재하기에, 특히 공개용 콘텐츠에 생성형 AI를 이용할 경우 저작권 침해가 일어나지 않도록 추가로 유의해야 합니다.
이용자로서의 법적 리스크보다 좀 더 넓은 상황도 볼까요.
생성형 AI는 저작물을 활용하여 구축되었을 뿐만 아니라, 창작노동자와 시장에서 경쟁함으로써 이들을 경제적으로 위협하기도 합니다. 생성형 AI가 창작자의 동의나 보상 없이 노동이 착취되는 구조에 기반하고 있다는 점에서 윤리적·정치경제적 문제를 던지기도 하는 대목입니다.
우리가 사용하는 생성형 AI의 데이터 수집과 생성 과정은 투명하며, 공정한 보상 체계가 자리잡고 있나요?

- **생성형 AI의 환경 비용**

생성형 AI는 환경적으로 비싼 기술입니다. 모델 학습 단계에서 대규모 컴퓨팅 자원이 소요되며, 이 과정에서 발생하는 탄소 배출량은 모델 하나당 수천 톤에 달하기도 합니다.
데이터센터의 냉각 시스템을 가동하기 위해 대량의 담수가 필요하고, GPU 등 하드웨어 생산·폐기 과정에서 희토류 채굴과 전자폐기물 문제가 수반됩니다. 데이터센터를 가동하기 위한 전력망 수요에 힘입어 화석 연료 및 원자력 발전이 힘을 얻고, AI 관련 실적을 내세우는 빅테크 기업들이 탄소배출 감축 등 ESG 목표에서 발을 빼는 사례도 생기고 있습니다.

일각에서는 환경비용이 과장되었다거나 기술 발전에 따라 개선될 것이라고 보기도 합니다. 설령 그렇다 해도 점점 많은 서비스에 생성형 AI가 활용되고 심지어 마이크로소프트 코파일럿처럼 운영체제 차원에서 상시, 수시로 생성형 AI를 구동하는 등 사용 규모 자체가 급증하는 추세를 감안하면 환경비용 문제를 가볍게 보기는 어렵습니다. 재생에너지 기반 데이터센터 전환이나 효율적 알고리즘 개발(경량화 등) 등 기술적 해결책이 모색되고 있으나, 기업의 자발적 노력에 의존하는 한계가 있습니다. 그런가 하면 더 많은 자원을 투입해서 AI 기술을 더 발전시키면 지구 온난화 등의 난제를 해결할 수 있고 지금의 환경 비용도 상쇄할 수 있을 것이라는 주장도 있지만, 과학적 전망보단 낭만적 기대에 가까운 것 같습니다.

생성형 AI의 간접적 환경 영향을 고려하고, 기후위기 시대에 환경적 책임을 다하려면 어떻게 해야 할까요? 현재로서는 환경비용을 파악하는 것조차 어렵습니다. 생성형 AI 구축 및 활용 과정에서의 탄소배출 등 환경비용은 기업 비밀 등의 사유로 잘 공개되지 않기 때문입니다. 이 정보를 공개하도록 요구하는 것에서 출발해볼 수 있겠습니다. 또한 AI 산업, 나아가 사회가 AI로 인해 본 이득을 기후위기 대응에 재투자하는 구조적 논의 역시 요청됩니다.

- **차별과 편향 문제**

생성형 AI 이전부터 각종 인공지능 및 자동화 시스템은 불투명한 방식으로 기존 편향을 재생산해 왔습니다. 과거 데이터를 학습하여 만들어진 생성형 AI 역시 사회 권력을 반영하는 편향을 재생산합니다. 예를 들어 특정 직군이나 문화적 맥락을 특정 성별, 인종, 계층과 결부하는 사회적 편향이 AI 생성 콘텐츠에서 표현될 수 있고 이는 채용, 콘텐츠 추천, 법률 판결 등 각종 분야에서 불공정한 결과로 이어질 수 있습니다. 원칙적으로 채용, 판결 등 사람에게 큰 영향을 미치는 중요한 의사결정에는 생성형 AI를 활용하지 않아야 할 것입니다. 소수자 집단을 대상화한 혐오표현이나 고정관념 또한 위험 요소가 됩니다.

- **공론장과 정보 생태계**

생성형 AI는 공론장과 정보 생태계에 위험으로 작용할 수 있습니다. 인간이 만든 것과 흡사해 보이는 글·그림 등을 자동적으로 생성하는 시스템의 보편화는 사회에 어떤 작용을 할까요?
우선 허위정보 생산에 필요한 비용이 생성형 AI로 인해 획기적으로 감소한다는 점을 생각해볼 수 있습니다. 텍스트, 이미지뿐만 아니라 상대적으로 고비용 매체인 동영상 역시도 '실제'와 생성물 사이의 구분이 어려워지고

있는데요. 악의적으로, 혹은 경제적 이윤을 위해 허위정보를 양산하는 것이 쉬워지는 만큼 '사실'의 위치는 좁아지고, 사실 검증에 들어가는 사회적 비용은 증가합니다.

확률 기반으로 작동하는 생성형 AI 시스템이 갖는 태생적 오류 가능성 또한 문제가 됩니다. 특히 자료 조사, 문서 작성 등 지식 관련 기능에 AI 시스템이 활용되는 추세는 곧 우리가 지식을 생산하고 습득하는 과정 전반에 해당 오류의 위험이 스며든다는 것을 뜻합니다.

정보를 소비하는 입장에서는 생성형 AI가 보편화할 수록 역설적으로 정확한 사실을 접하기 위한 비용이 늘어날 수도 있으며, 메시지를 발신하는 입장에서는 값싸게 생산되는 (허위·저품질일 수 있는) 정보와 사람들의 관심을 두고 경쟁하게 되는 문제가 생길 수 있습니다.

- **딥페이크**

생성형 AI 기술의 대표적인 오용 사례로는 딥페이크를 꼽을 수 있습니다. 딥페이크는 누군가가 하지 않은 말이나 행동을 한 것처럼 묘사하는 합성물로, 특히 개인을 상대로 한 성착취 등 폭력이나 사기 등에 활용될 위험이 큽니다. 이미 한국에서 조직적인 딥페이크 성범죄가 큰 사회적 문제로 대두된 바 있습니다. 이러한 범죄는 생성형 AI

기술 이전에도 존재했지만 생성형 AI 기술은 해당 범죄를 훨씬 쉽게 만들며 처벌과 예방, 기술적 대응, 피해자 회복 등 각종 영역에서 새로운 문제를 발생시킵니다.

- **보안과 프라이버시**

생성형 AI 모델의 성능은 학습 데이터의 양과 (매개변수의 개수로 표현되는) 모델 크기가 커질 수록 증가하는 경향을 보여왔습니다. 그렇기 때문에 생성형 AI 산업은 최대한 많은 데이터의 수집을 추구하며, 그 과정에서 데이터의 적법성이나 품질 관리는 상대적으로 등한시되어 왔습니다. 온라인에 공개된 각종 개인정보를 수집하여 생성형 AI 구축에 활용하는 행태는 프라이버시에 대한 위협을 제기하는 한편, 개인정보 수집을 최소화해야 한다는 개인정보보호 원칙에도 위배될 소지가 있습니다. 나아가 이렇게 수집된 정보는 생성형 모델의 출력을 통해 다른 사람에게 노출될 위험이 있습니다.

모델 구축(학습) 단계뿐만 아니라 활용 단계에서도 데이터 수집이 이루어질 수 있습니다. 챗지피티 등의 서비스에 입력하는 질문 내역 등이 대표적인 예입니다. 특히 이런 경우 개인정보뿐만 아니라 민감한 업무 자료 등 역시도 보안 위협의 대상이 됩니다. 또한 윈도우 OS에 탑재된 코파일럿(Copilot)이나 메타의 AI 안경 등의 사례에서 볼 수 있는 것처럼, 이같은 데이터 수집이 이루어지는 지점은

특정 웹 서비스를 넘어 사용자의 컴퓨팅 환경 전반과
일상 공간으로 확장하는 추세라는 점에서 보안 위험이
발생할 수 있는 지점 또한 늘어난다고 볼 수 있습니다.

3장. 시민사회 생성형 AI 정책 모델

[단체명] 생성형 AI 활용 정책

1. 총칙

1) 목적

이 정책은 우리 단체가 생성형 AI(Generative AI) 기술을
단체의 활동 목적과 인권 원칙을 준수하면서 책임감 있고 효과적으로
사용하기 위한 기준과 절차를 정하는 것을 목적으로 한다.

2) 기본 원칙

우리는 생성형 AI 기술을 사용할 때 다음과 같은 원칙을 준수한다.

① 생성형 AI를 사용하여 만들어진 결과물과 의사결정에
대한 모든 책임은 우리 단체에게 있다.
② 생성형 AI는 보조적인 도구일 뿐이며, 활동가의
판단과 숙련을 대체하지 않는다.
③ 생성형 AI의 결과물은 소수자 및 취약계층에 대한 어떠한 형태의 편향이나
차별을 포함하거나, 기본권에 부정적인 영향을 미치지 않아야 한다.
④ 생성형 AI 활용 과정에서 개인정보와 보안이 침해되지 않아야 한다.
⑤ 생성형 AI가 결과물 생성에 핵심적으로 기여했거나 혼동을 야기할 우려가
있는 경우, 생성형 AI의 활용 여부 및 방식을 투명하게 공개한다.
⑥ 생성형 AI 기술의 발전이 환경과 노동에 미치는 영향을 고려한다.

3) 이 정책의 범위

이 정책은 우리 단체가 외부의 상용 생성형 AI 서비스를 이용하는 경우를
대상으로 한다. 단체가 AI 도구를 직접 개발 제공하거나 생성형 AI가 아닌
다른 종류의 AI 도구를 이용하는 경우에는 별도의 원칙과 지침을 마련한다.

2. 생성형 AI 활용 지침

1) 정보의 정확성 확인

생성형 AI의 결과물은 부정확한 정보를 포함할 수 있으므로,
반드시 신뢰할 수 있는 방식으로 정확성을 확인해야 한다.

① 사실관계가 중요한 업무에서는 생성형 AI 활용에 특별한 주의가 필요하다
② 인터넷 검색, 전문가 자문 등 다양한 출처를 통해 사실관계를 교차검증한다.
③ 데이터나 자료가 최신 정보인지 확인한다.
④ 권위있는 출처나 공식 문서를 우선적으로 참조한다.
⑤ 최근 정보를 반영한 AI의 결과물(예 : 웹검색 기반
AI 결과물)을 우선적으로 활용한다.
⑥ 명확하고 구조화된 프롬프트를 사용하고 출처를 제공할 것을 요구한다.
⑦ 원본을 직접 읽지 않은 상태에서 요약 기능만 사용하는 것은 주의해야 한다.

2) 편향 및 고정관념(stereotype)에 대한 비판적 검토

AI는 기존의 데이터를 학습하고 이를 모방하기 때문에,
생성형 AI의 결과물이 기존 현실의 편견, 편향, 고정관념을 그대로 반영할 수
있으므로, 이러한 결과물이 공개되거나 사용되지 않도록 주의해야 한다.

① 활동가가 생성형 AI 결과물의 편향·차별 표현을 인식할 수 있도록 정기적인
인권 교육을 시행한다. [또는 AI 결과물 검토 담당자를 지정한다]
② 생성형 AI 활용 과정에서 문제가 될 수 있는 표현이 발견되면,
즉시 사용을 중지하고 [검토 담당자]에게 전달한다.
③ 생성형 AI에 차별적이지 않은 내용으로 다시 작성해줄
것을 요구하고, 그 결과물을 다시 검토한다.
④ 해당 생성형 AI를 서비스하는 업체에 발생한 문제에 대해 신고한다.
⑤ 차별적, 혐오적 표현을 반복적으로 생성할 경우,
해당 생성형 AI의 사용을 중단한다.
⑥ 생성형 AI에만 의존하지 않고 다른 경로를 통한 정보와 관점의 수집을 검토한다.

3) 개인정보 보호 및 보안

상용 생성형 AI 서비스를 이용할 경우, 프롬프트로 입력한 데이터는 AI 업체의 서버에 저장되므로 무단 접근 및 유출과 같은 보안 문제가 발생할 수 있다. 또한, 이 데이터가 모델 재학습에 사용될 경우 다른 이용자의 출력물을 통해 개인정보나 기밀정보가 유출될 가능성도 있다. 이에 정당한 법적 근거없이 개인정보가 처리되거나 단체의 기밀 정보가 유출되지 않도록 보안에 주의해야 한다.

① 프롬프트를 통해 주민등록번호, 신용카드번호, 비밀번호, 민감정보(생체정보, 성적지향 등)와 같은 개인정보를 올리지 않는다.
② 생성형 AI를 통한 개인정보의 분석이 필요할 경우 가명처리를 해야 한다.
③ 보안 등급에 따라 높은 수준의 보안이 필요한 기밀자료(예 : 피해자 인터뷰, 비공개 회의록, 회계기록 등)를 프롬프트를 통해 올리지 않는다.
④ 생성형 AI 서비스의 이용약관, 개인정보 처리방침, 보안 정책을 확인하여, 데이터 보관 기간, 프롬프트로 입력된 데이터를 AI 학습에 사용하는지 여부, 개인정보보호법 등 관련 법 준수 여부, 암호화 등 보안정책, 요금제별 보안 수준의 차이를 파악한다. 가능하면 학습 데이터 활용을 거부(옵트아웃)하는 옵션이나 해당 기능이 포함된 요금제를 선택한다.
⑤ 생성형 AI를 통해 공유한 데이터는 정기적으로 백업하고 삭제한다.
⑥ 생성형 AI가 다른 애플리케이션이나 외부의 API와 연계될 경우, 그 과정에서 불필요한 개인정보나 데이터가 전송되지 않도록, 전송되는 데이터의 범위를 확인한다.
⑦ 업무용 계정과 개인용 계정을 분리하여 사용한다.

4) 저작권

생성형 AI 사용시 여러 측면에서 저작권 침해 위험이 있다. 사회적으로는 저작권자의 데이터를 허락없이 AI 학습 데이터로 사용할 수 있는지가 논란이 되고 있지만, 이는 이용자가 통제할 수 없는 부분이다. 다만, 학습에 사용된 개인정보나 저작물이 모델에 암기되어 결과물에 반영될 수 있기 때문에, 생성형 AI의 결과물이 학습에 사용된 저작물과 상당히 유사하게 생성될 경우, 이용자의 의도와 무관하게 저작권 침해 책임을 질 수 있다.

① 생성형 AI의 결과물(특히, 이미지나 오디오)이 의도하지
않게 저작권을 침해할 수 있으므로 주의한다. 사용하기 전에
유사한 저작물이 있는지 검색(예 : 이미지 검색)해본다.
② 생성형 AI 결과물을 재료로 사람이 직접 상당한 수정·편집을 거쳐 활용한다.

5) 생성형 AI 활용의 투명성

생성형 AI를 사용하는 사실을 수용자들이 인지하지 못해 오해나 혼란을 야기할 우려가 있을 경우, 해당 결과물이 생성형 AI를 통해 만들어졌음을 표시한다.

① 생성형 AI를 활용한 분석, 또는 생성형 AI로 제작한 음악, 이미지, 영상 등
생성형 AI가 결과물 생성에 핵심적으로 기여한 경우, 해당 저작물이
생성형 AI에 의해 만들어졌다는 사실을 표시한다.
② 딥페이크와 같이 생성형 AI를 이용해 현실과 혼동될 수 있는 결과물을
생성한 경우, 그 사실을 저작물에 표시한다. 다만, 예술적·창작적
저작물의 경우, 감상을 저해하지 않는 방식으로 표시할 수 있다.
③ 챗봇, 동시통역 도구 등 외부 사람과 직접적으로 상호작용하는
생성형 AI의 경우, 사람들이 자신과 상호작용하고 있는 대상이
AI라는 점을 명확하게 인지할 수 있도록 알린다.
④ 홈페이지 등을 통해 본 단체의 생성형 AI 정책을 공개한다.

6) AI가 환경에 미치는 영향에 대한 고려

생성형 AI의 확대에 따라 데이터센터 운영을 위한 전력사용량과 물 사용의 증가, AI를 위한 반도체 생산 등 자원사용량이 증가하고 있다. 이에 환경에 미치는 부정적 영향을 최소화할 수 있는 방식으로 AI를 활용한다.

① 감사인사 등 불필요한 대화 또는 에너지를 많이 사용하는
이미지·음성·영상 처리 요청을 지양한다.
② 동일한 자료를 재요청하는 경우가 많을 때는 생성된 결과물의 재활용,
단체 구성원간 결과물 공유 등을 통해 불필요한 반복 요청을 최소화한다.

③ 생성형 AI가 아니어도 처리가 가능한 업무는 적합한 다른 도구를 우선 활용한다.
④ 가능하다면 경량 AI 모델을 사용한다.
⑤ AI 운영을 위한 데이터센터의 환경영향평가, 전력사용량, 에너지효율 등의 정보공개, 친환경 재생에너지 사용 등 친환경 정책을 실천하는 기업의 제품을 사용한다.

3. 정책의 수립과 집행

1) 생성형 AI 사용 승인

① 단체의 사업 목적으로 생성형 AI를 사용하기 위해서는 [운영위원회]의 승인을 거쳐야 한다.
② 특정 생성형 AI 사용을 승인하기 전에, 해당 AI 서비스의 성능, 적절한 요금제, 필요한 설정 등 사용 정책을 마련한다.
③ AI 책임자는 본 단체에서 활용하는 생성형 AI의 목록을 관리하고, 변경시 구성원에게 공지한다.
④ 생성형 AI로 인해 업무를 대체하거나 변경할 필요가 있을 경우, 단체의 구성원과 사전에 협의한다.

2) 생성형 AI 활용이 허용되는 활용의 범위

AI 책임자는 본 단체에서 생성형 AI의 활용이 허용되는 사례, 허용되지 않는 사례, 또는 엄격한 검토가 필요한 활용 사례를 문서로 관리한다.

3) 교육 및 역량 강화

① 모든 구성원이 본 정책을 숙지하고, AI 관련 최신 동향에 대해 인지할 수 있도록 [1년에 1회 이상] 구성원에 대한 AI 교육을 시행한다.
② 업무상 필요한 도구의 사용법에 대한 교육의 일환으로, 생성형 AI의 사용법에 대한 교육을 시행한다.
③ 단체 구성원의 역량 강화를 위해 필요할 경우, 업무 수행 과정에서 생성형 AI의 도움을 받는 것을 제한할 수 있다.

4) 외부 파트너와의 협력

다른 단체 또는 외부 사람들과 협업을 하거나, 기고를 받는 등 단체의 활동을 위해 협력할 때, 생성형 AI 활용 정책에 대해 외부 사람에게 사전에 고지하거나, 해당 정책에 대해 협의해야 한다.

5) 문제발생 시 조치

① 생성형 AI와 관련되어 문제가 발생할 경우 즉시 'AI 책임자'에게 보고한다. 보고에는 다음과 같은 내용을 포함한다 : 발생 일시 / 사용 도구명 / 해당 결과물 / 문제가 된 부분 / 프롬프트 입력 내용 / 부정적 영향의 내용과 범위
② AI 책임자는 즉시 사실을 확인하고, 필요할 경우 피해 확산을 방지하기 위한 긴급조치를 취한다.
③ AI 책임자는 [운영위원회]를 소집하여 단체의 대응 방안을 수립한다. 이를 위해 문제의 원인, 영향의 범위, 단체의 책임 유무 및 범위, 관련 법제 및 법적 대응의 필요성 등을 검토한다.
④ 필요할 경우 적절한 방식으로 해당 사안에 대해 외부에 공지한다. 공지에는 문제의 내용 및 원인, 영향을 받는 당사자, 단체의 대응 조치, 재발 방지 조치 등이 포함될 수 있다.
⑤ 필요할 경우 적절한 방식으로 영향을 받는 당사자에게 사과문을 전달한다. 사과문에는 문제의 내용 및 원인, 단체의 대응 조치, 피해 구제 및 보상, 재발 방지 조치 등이 포함될 수 있다.
⑥ 재발 방지 대책을 수립하고 필요하다면 본 정책에 반영한다
⑦ AI 책임자는 본 사안에 관련된 모든 내용과 과정을 기록한다.

6) AI 책임자와 감독

① 본 단체의 책임있는 AI의 활용과 감독을 위해 'AI 책임자'를 둔다. 본 단체의 AI 책임자는 [　　] 이다.
② 생성형 AI의 결과물이 본 단체의 정책에 부합하지 않거나, 본 정책을 위반하는 사례가 있을 경우 AI 책임자에게 보고한다.
③ 단체 구성원이 본 정책을 위반할 경우 내부 징계 절차에 따른다.

7) 정책의 변경

① AI 기술의 급속한 발전을 고려하여 본 정책은 AI 책임자가 필요하다고 판단할 경우, 또는 최소한 연 1회 재검토 및 업데이트 되어야 한다.
② AI가 단체에 미치는 영향에 대해 정기적으로 평가한다.
③ 본 정책에 대해 논의할 때 모든 구성원이 참여할 수 있도록 한다.

4장. 시민사회의 생성형 AI 정책 모델 해설

1. 시민사회 생성형 AI 정책 모델의 개요

생성형 AI는 시민사회단체의 업무 효율을 높이고, 기존의 활동 방식을 변화시킬 수 있는 도구가 될 수 있습니다. 그러나 동시에 부정확한 정보 생성, 결과물의 편향성, 개인정보 및 단체 기밀의 유출, 기술 의존으로 인한 활동가 역량 약화 등 다양한 위험도 포함하고 있습니다. 이러한 문제는 단체의 사회적 책임, 신뢰도, 그리고 인권 문제와 직결됩니다.

시민사회단체는 공익, 인권, 투명성, 민주성을 핵심 가치로 삼고 있습니다. 따라서 생성형 AI를 활용할 때도 단체의 목적과 가치에 맞는 명확한 기준과 절차, 그리고 책임 구조가 필요합니다. 이것이 바로 시민사회단체의 생성형 AI 정책이 필요한 이유입니다.

생성형 AI를 활용하는 방식은 단체의 성격, 규모, 활동 분야에 따라 매우 다릅니다. 같은 단체 안에서도 활동가의 역할에 따라 주로 사용하는 AI 도구나 활용 정도는 달라질 수 있습니다. 따라서 생성형 AI 정책은 모든 단체에 동일하게 적용되는 정답이 존재하는 것이 아니라, 각 단체가 스스로 토론하고 결정해야 할 문제입니다.

이 정책 모델은 모든 단체가 따라야 하는 규범을 제시하는 것이 아니라, 각 단체가 자신의 상황과 철학에 맞는 정책을 설계할 수 있도록 돕는 기본 틀을 제공하는데 목적이 있습니다. 물론 이 정책 모델은 생성형 AI를 사용할 때 고려해야 할 기본 원칙도 함께 제시합니다.

단체는 이를 참고해 단체의 상황에 맞게 조항을 수정·삭제·추가하여 내부 정책을 수립할 수 있습니다. 또한, 이 가이드와 정책 모델이 생성형 AI 사용을 권장하거나 장려하는 것으로 오해되어서는 안됩니다. 생성형 AI가 충분한 효율성을 제공하지 않는 경우, 환경적 부담에 대한 우려, 또는 기술에 대한 불편함 등 다양한 이유로 생성형 AI를 사용하지 않을 단체나 활동가가 있을 수 있습니다. 이 가이드의 목적은 어디까지나 "만약 단체가 생성형 AI를 사용할 경우, 어떤 최소한의 기준 하에서 사용하는 것이 바람직한가"를 제안하는 것입니다.

시민사회는 단순한 AI 사용자를 넘어, 신뢰할 수 있는 AI의 개발과 책임있는 사용을 요구하는 역할을 합니다. 따라서 시민사회의 AI 정책은 단체 내부의 실무지침이면서 동시에 사회적 정책 제안의 의미도 갖습니다. 각 단체가 이 정책 모델을 바탕으로 자신들의 상황에 맞게 발전시키고 이를 실천해 나간다면 책임있는 AI 활용 문화 형성에 기여할 수 있을 것입니다.

2. 총칙

정책 모델의 형식은 각 단체가 선호하는 방식으로 자유롭게 구성할 수 있습니다. 예를 들어, 법령이나 약관과 유사하게 '1장 총칙', '1조(목적)'과 같은 형식을 사용할 수도 있습니다.

1) 목적

> 이 정책은 우리 단체가 생성형 AI(Generative AI) 기술을 단체의 활동 목적과 인권 원칙을 준수하면서 책임감있고 효과적으로 사용하기 위한 기준과 절차를 정하는 것을 목적으로 한다.

생성형 AI 정책을 수립하는 핵심 목적은 단체가 이 기술을 단체의 가치와 인권 원칙에 부합하도록 사용하기 위함입니다. 여기서 '책임감'있는 사용은 단순히 도구를 효율적으로 사용하는 것을 넘어, 생성형 AI 사용이 미칠 수 있는 사회적 영향, 예를 들어 편향과 차별, 노동 및 환경에 미치는 영향을 무시하지 않고 고려하겠다는 의미입니다. '효과적'인 사용 역시 업무 효율성만을 뜻하지 않습니다. 생성형 AI의 사용이 단기적으로 효율적인 것처럼 보이더라도 활동가의 역량을 훼손하거나 조직 내 숙고·토론의 과정을 대체한다면, 그것은 효과적인 사용이 아닐 것입니다. 사실 확인을 하는데 더 많은

시간이 들거나, 귀찮다는 이유로 결과물을 제대로 검토하지 않아 잘못된 결정을 하거나 단체의 명성을 훼손하게 된다면, 이 역시 효과적인 사용은 아닐 것입니다. 각 단체는 자신의 상황과 요구에 맞는 적절한 활용 방법을 고민할 필요가 있으며, 이 정책은 그러한 고민의 결과를 반영해야 할 것입니다.

2) 기본 원칙

> 우리는 생성형 AI 기술을 사용할 때 다음과 같은 원칙을 준수한다.
>
> ① 생성형 AI를 사용하여 만들어진 결과물과 의사결정에 대한 모든 책임은 우리 단체에게 있다.
> ② 생성형 AI는 보조적인 도구일 뿐이며, 활동가의 판단과 숙련을 대체하지 않는다.
> ③ 생성형 AI의 결과물은 소수자 및 취약계층에 대한 어떠한 형태의 편향이나 차별을 포함하거나, 기본권에 부정적인 영향을 미치지 않아야 한다.
> ④ 생성형 AI 활용 과정에서 개인정보와 보안이 침해되지 않아야 한다.
> ⑤ 생성형 AI가 결과물 생성에 핵심적으로 기여했거나 혼동을 야기할 우려가 있는 경우, 생성형 AI의 활용 여부 및 방식을 투명하게 공개한다.
> ⑥ 생성형 AI 기술의 발전이 환경과 노동에 미치는 영향을 고려한다.

생성형 AI 정책이 일관성을 가지려면,
그 기반이 되는 원칙을 명확히 할 필요가 있습니다.
본 정책 모델은 6가지 원칙을 제안하고 있습니다.

첫째, 생성형 AI를 사용하여 만들어진 결과물과 그에 기반한 의사결정에 대한 모든 책임은 우리 단체에게 있습니다.
아무리 AI가 일정한 자율성을 가진다해도, 도구에게 책임을 물을 수는 없습니다. 생성형 AI 도구의 기능상 결함 때문에 문제가 발생하여 향후 AI 개발업체에 문제 제기를 할 수는 있겠지만, 1차적인 책임은 그 결과물을 사용한 우리 단체에게 있습니다. 따라서 각 단체는 생성형 AI를 사용하는 전 과정에서, AI 활용의 주체로서 책임을 다하기 위한 절차를 마련해야 합니다. 예를 들어, 생성형 AI의 결과물은 항상 단체가 책임지고 검토를 해야 하며, 문제가 발생했을 때 어떻게 대응할 것인지에 대한 내부 절차를 마련해야 합니다. 또한, 이 원칙은 생성형 AI를 사용하는 모든 과정에서, 설사 생성형 AI의 결과물을 거의 그대로 사용하더라도, 최종적인 판단과 감독은 인간(단체 활동가)의 책임 아래 이루어져야 한다는 것을 의미합니다.

둘째, 생성형 AI는 어디까지나 보조적인 도구일 뿐이며, 활동가의 판단과 숙련을 대체해서는 안됩니다.
이는 첫번째 원칙과도 연결됩니다. 생성형 AI의 결과물을 1차적으로 판단하는 것은 단체 활동가입니다. 생성형 AI는 활동가를 대체하는 것이 아니라, 거꾸로 활동가의

역량을 강화하는 도구가 되어야 합니다. 이를 위해서는 단체 활동가가 AI를 책임 있고 효과적으로 사용할 수 있는 역량을 갖추고 있어야 합니다. 단체는 활동가가 전문성, 경험, 역량을 키워나갈 수 있도록 지원해야 합니다. 이 정책 모델은 신입 활동가의 역량 강화를 위해 문서 작성에 생성형 AI의 사용을 일정하게 제한하는 것과 같이, 필요하다면 업무 수행 과정에서 생성형 AI의 도움을 받는 것을 제한할 수 있도록 제안합니다. 이는 개인 활동가의 역량 문제만이 아니라, 단체 내에서 특정 이슈에 대한 입장을 토론하고 공통의 문제의식을 유지하는 것은 필수적이며, 생성형 AI가 이를 대체하도록 해서는 안되기 때문입니다. 단체의 성명이나 입장을 작성할 때 생성형 AI를 사용하면, 단체 내부 논의와 숙고의 과정을 약화시키거나 대체할 우려가 있으므로, 단체에 따라 이러한 용도의 사용을 금지하는 정책을 둘 수도 있습니다.

셋째, 생성형 AI의 결과물이 소수자 및 취약계층에 대한 어떠한 형태의 편향이나 차별을 포함하거나 기본권에 부정적인 영향을 미쳐서는 안됩니다. 생성형 AI 학습에 사용된 데이터는 현실의 편향이나 불평등, 고정관념을 그대로 반영하고 있으며, 생성형 AI 결과물은 이를 재생산할 수 있습니다. 편향되거나 차별적인 결과물을 사용하는 것은 소수자 및 취약계층에게 2차 피해를 야기하고, 동시에 인권 옹호를 지향하는 단체의 신뢰와 명성을 훼손할 수 있습니다. 〈2-2) 편향 및 고정관념(stereotype)에 대한 비판적 검토〉에서

이러한 위험을 줄이기 위한 지침을 다루고 있습니다.

넷째, 생성형 AI 활용 과정에서 개인정보와 보안이 침해되지 않아야 합니다. 개인정보 및 보안 이슈는 모든 디지털 활동에서 중요한데, 생성형 AI를 사용할 때에도 예외가 아닙니다. 특히, 외부 상용 생성형 AI 서비스를 이용할 경우, 프롬프트를 통해 입력한 데이터가 해당 업체로 전송될 수밖에 없으며 그에 따른 보안 문제가 발생합니다. 또한, 이렇게 AI 업체에 제공된 데이터는 향후 AI 학습 목적으로 사용될 수 있으며, 그 결과 AI 제품이 활용되는 과정에서 다른 사용자의 결과물로 출력될 수 있습니다. 〈2-3) 개인정보 보호 및 보안〉에서 이에 대비한 구체적인 지침을 다루고 있습니다. 각 단체의 개인정보 및 보안 정책 역시 생성형 AI를 고려하여 업데이트될 필요가 있습니다.

다섯째, 생성형 AI가 결과물 생성에 핵심적으로 기여했거나 혼란을 야기할 우려가 있는 경우, 생성형 AI의 활용 여부 및 방식을 투명하게 공개할 필요가 있습니다. 인공지능의 맥락에서 투명성과 설명가능성은 다양한 의미를 포함합니다. 우선 사람들이 자신과 상호작용하는 대상이 AI임을 인지할 수 있도록 해야 합니다. 또한 AI 시스템의 결정이 추적 가능하고 설명 가능해야 함을 의미합니다. 즉, 문제가 생겼을 때 그 원인을 추적할 수 있어야 하고, AI 시스템이 내린 결정의 근거, 논리, 영향을 미친 주요 요소 등에 대해 알 수 있어야 합니다. 더불어, AI 시스템의 개발자는 이용자에게,

AI 시스템의 이용자는 그 영향을 받는 사람에게 필요한 정보를 제공할 필요가 있습니다. 물론 이러한 원칙이 생성형 AI 서비스를 사용하는 이용자에게 동일하게 적용되지는 않을 수 있습니다. 생성형 AI의 경우 결과물을 도출한 근거나 논리가 결과물 자체에 포함되어 있거나 또는 중요하지 않을 수 있기 때문입니다. (예를 들어, 왜 이러한 이미지를 출력했는지는 프롬프트를 입력한 이용자가 바로 알 수 있습니다. 반면 어떤 학습 데이터로부터 그러한 이미지가 출력되었는지 설명하는 것은 불가능할 수 있습니다.)

기본적으로 책임성과 투명성 원칙을 중요하게 생각하는 시민사회단체는 생성형 AI의 활용과 관련해서도 가능한 범위에서 투명성을 확보할 필요가 있습니다. 그 이유는 생성형 AI의 결과물에 영향을 받는 사람들이, 정보에 기반하여 판단할 수 있도록 하여 AI와 단체에 대한 신뢰성을 높일 수 있기 때문입니다. 예를 들어, 생성형 AI를 이용해 어떤 문서를 요약했으나 그 정확성이 완전하지 않을 수 있다면, 해당 결과물에 그 사실을 표시함으로써 수용자들이 이를 감안하여 정보의 신뢰도를 판단할 수 있도록 해야 합니다. 또한, 딥페이크처럼, 수용자가 생성형 AI의 결과물을 인간의 창작물 또는 실재로 오인함으로써 혼란이 발생할 수 있습니다. 예를 들어, 미국 국방부 청사(펜타곤) 부근에서 대형 폭발이 발생한 것처럼 보이는 가짜 사진이 소셜미디어에서 확산되어 혼란이 발생한 바 있습니다.

경우에 따라 투명성은 법적으로 요구되는 의무일 수 있습니다. 예를 들어, EU AI 법에 따르면, 1) AI 시스템의 제공자는 사람이 자신과 상호작용하는 대상이 AI라는 것을 알 수 있도록 설계해야 하며, 2) 생성형 AI의 경우 제공자는 결과물이 AI 생성물임을 기계판독 가능한 방식으로 인식될 수 있도록 해야 합니다. 또한, 3) 감정인식이나 생체인식 시스템의 배치자(이용자)는 자연인에게 그 사실을 알려야 하며, 4) 딥페이크를 생성하는 AI 시스템 배치자(이용자)는 결과물이 AI에 의해서 생성되었다는 사실을 공개해야 합니다. 단, 예술적 작품에 해당할 경우 해당 작품의 감상을 저해하지 않는 방식으로 알릴 수 있습니다. 시민사회단체의 경우 4번째 의무와 관련이 많을 것입니다. 단체에서 권력을 비판하는 패러디 이미지를 만들거나 특정 이슈와 관련된 다큐멘터리를 만드는 경우가 많기 때문입니다.

한국의 인공지능 기본법도 인공지능 투명성 확보 의무(31조)를 규정하고 있습니다. 1) 고영향 또는 생성형 인공지능에 기반하여 운영된다는 사실을 이용자에게 사전에 고지해야 하고, 2) 생성형 인공지능의 결과물이 그것에 의해 생성되었다는 사실을 표시해야 하며, 3) 생성형 AI를 이용해 딥페이크를 제작하는 경우 그 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지 또는 표시해야 합니다. 이때 예술적·창의적 표현물인 경우 전시 또는 향유를 저해하지 않는 방식으로 할 수 있습니다. 한국의 인공지능 기본법에서 이러한 의무의 주체는

인공지능 사업자입니다. 그렇기 때문에, 생성형 AI 도구의 이용자인 시민사회단체가 이 의무의 주체는 아닐 수 있습니다. 다만, 법제가 아직 형성 중인 단계에 있어 해석이 유동적이고, 개정 가능성이 크며, 투명성 의무의 취지를 고려할 때, 신뢰와 인권을 중요시하는 시민사회단체의 입장에서 가능한 범위에서 투명성 원칙을 자율적으로 준수하는 것이 바람직할 것입니다.

그러나 생성형 AI 사용 사실을 모든 결과물에 일괄적으로 표시하는 것은 비현실적이며 필요 이상으로 부담이 될 수 있습니다. 인터넷 검색이나 사무용 애플리케이션 등에 AI 기능이 기본 내장됨에 따라, 이용자의 의도와 상관없이, 정도의 차이는 있을지언정 상당히 많은 업무에 AI를 활용하게 되는 상황이 발생하고 있습니다. 이때 AI의 도움을 받은 모든 결과물에 '이 결과물의 생성에 AI의 도움을 받았습니다'라고 표시를 하는 것은 단체에 실무적인 부담이 될 뿐만 아니라, 수용자에게도 별다른 정보를 제공하지 못합니다. 단체가 충분한 검토를 거쳐 책임 있게 공개한 결과물에 AI 사용 사실을 기계적으로 표기할 경우, 오히려 그 결과물에 대한 대중의 신뢰를 불필요하게 떨어뜨릴 수도 있습니다. 첫번째 원칙에서처럼 담당자 및 단체가 모든 표현이나 사실관계를 엄격하게 검토한 후, 공개되는 결과물에 대한 모든 책임을 질 수 있다면, 생성형 AI는 다른 도구와 마찬가지로 하나의 도구로 활용한 것으로 볼 수 있습니다. 그러나 사람의 검토에도 불구하고 결과물의 핵심적인 내용에 있어서

AI의 기여를 대체할 수 없는 경우, 예를 들어 AI 도구를 활용하여 데이터 분석 결과를 도출했거나, 생성형 AI로 음악, 이미지, 영상 등의 저작물을 제작한 경우, 생성형 AI 사용 여부 및 방식(어떤 방식으로 사용했는지 또는 생성형 AI가 결과물에 어떻게 기여했는지)을 표시하는 것이 바람직할 것입니다. 특히 딥페이크와 같이 현실과 혼동을 야기할 수 있는 결과물의 경우에는 수용자의 혼란을 방지하기 위해서라도 생성형 AI 이용 사실을 표시할 필요가 있습니다.

투명성 원칙을 어떻게 적용할 것인지는 이 가이드와 정책 모델의 논의 과정에서도 매우 논쟁적이었던 주제입니다. '생성형 AI가 결과물 생성에 핵심적으로 기여했거나 혼동을 야기할 우려가 있는 경우'라는 기준이 모호하기 때문에, 자의적인 판단에 따라 생성형 AI의 활용 여부를 공개하지 않을 수 있다는 우려도 제기되었습니다. 그러나 이 가이드의 제안은 법적인 판단을 하거나 객관적인 기준을 제시하기 위한 것이 아니라는 점을 다시 한번 확인하고자 합니다. 어떠한 기준으로 생성형 AI 활용 여부에 대해 공개할 것인지는 단체의 윤리 기준과 토론의 결과를 반영하여 결정되어야 할 것입니다. 〈2-5) 생성형 AI 활용의 투명성〉에서 투명성에 관한 지침을 다루고 있습니다.

여섯째, 생성형 AI 기술의 발전이 환경과 노동에 미치는 영향을 고려해야 합니다. AI는 학습 및 운영 과정에서 전기와 물과 같은 많은 자원을 소비합니다. AI 학습 및 운영에 대규모

연산이 필요하기 때문인데 AI의 성능 향상을 위해 학습에 사용되는 데이터의 규모와 매개변수의 크기도 증가하고 있으며, 이에 비례하여 AI의 에너지 수요도 급증하고 있습니다. 현재 에너지 공급의 상당 부분은 여전히 석탄이나 천연가스와 같은 고탄소 에너지원에 의존하고 있기 때문에 AI와 데이터센터의 확장이 기후 위기를 악화시킨다는 우려가 커지고 있습니다. 또한, 데이터센터에서 발생하는 열을 냉각하기 위해 많은 양의 물이 소비되기 때문에 지역사회와의 갈등이 일어나기도 합니다. 환경단체가 아니더라도 기후 위기 대응에 공감하는 시민사회단체라면 이러한 문제를 외면할 수 없습니다. 물론 AI 학습 및 운영 과정의 에너지 문제는 소비자인 개별 시민사회단체가 직접 개입하기 어려운 구조적 문제입니다. 그럼에도 불구하고, 유사한 기능을 제공하면서도 에너지를 덜 소비하는 경량 모델을 사용하려 하고, AI 업체에 그러한 모델의 제공을 요구할 수 있습니다. AI 업체들이 AI 개발 및 운영 과정에서 얼마나 많은 에너지를 사용하는지에 대한 데이터를 투명하게 공개할 것을 요구할 수도 있습니다.

한편, AI의 발전으로 기존의 직무가 대체될 수 있다는 우려도 커지고 있습니다. 생성형 AI 역시 예외가 아닙니다. 이미 프로그래머, 통역사, 디자이너, 콜센터 상담원 등 노동자들이 생성형 AI의 도입으로 인해 해고되거나 일자리가 줄어드는 현상이 나타나고 있습니다. 이는 물론 국가, 사회적인 차원에서 다뤄야 할 구조적인 문제이지만, 개별 단체 차원에서도

스스로의 책임을 고민해야 할 영역입니다. 단체가 생성형 AI를 도입할 때 기존에 업무를 수행하던 활동가와 함께 협의하는 과정이 필요합니다. 생성형 AI가 예상보다 기존의 업무를 대체하는데 한계가 있을 수도 있고, 일정하게 대체할 경우 기존의 직무를 조정해야 할 수도 있습니다. 재정이 열악한 시민사회단체 입장에서 생성형 AI 기술은 기존에 비용을 감당할 수 없었던 일을 가능하게 하거나 비용을 절감할 수 있는 수단이 될 수 있습니다. 그러나 두번째 원칙에서 언급한 것처럼 생성형 AI에 의존하는 것이 단기적으로 효율적이더라도 활동가 및 단체의 전문성이나 역량을 유지하고 강화하는데 도움이 되는 것인지 신중하게 판단할 필요가 있습니다.

3) 이 정책의 범위

> 이 정책은 우리 단체가 외부의 상용 생성형 AI 서비스를 이용하는 경우를 대상으로 한다. 단체가 AI 도구를 직접 개발 제공하거나 생성형 AI가 아닌 다른 종류의 AI 도구를 이용하는 경우에는 별도의 원칙과 지침을 마련한다.

이 정책은 챗지피티(ChatGPT)나 제미나이(Gemini), 클로드(Claude) 등 상용 생성형 AI 서비스를 단체가 활용하는 경우에 초점을 맞춘 것입니다. 그러나 단체가 AI를 활용하는 방식은 매우 다양할 수 있기 때문에, 모든 사례에 이 정책을

그대로 적용하는 것이 적절하지 않을 수 있습니다. 예를 들어, 단체 홈페이지에 방문자용 챗봇을 설치하여 정보를 제공하거나 문의에 응답할 수 있도록 하는 경우, 국제회의나 행사에서 AI 동시통역 서비스를 사용하는 등의 경우, 본 정책을 그대로 적용하기 어렵습니다. 동시통역 AI 서비스의 통역 결과물에 할루시네이션이 포함되더라도 이에 실시간으로 대응하기는 곤란할 것이기 때문입니다. 다만, 이러한 서비스를 도입할 것인지 고려하는 과정에서, 본 정책의 원칙과 지침에 기반하여 사전에 엄격하게 평가할 수 있을 것입니다. 또한 오픈소스 모델을 이용하여 단체가 자체적으로 생성형 AI를 구축하여 사용하는 경우에도 이 정책을 적용할 수 있겠지만, 이 경우 생성형 AI 개발 과정에 대한 정책을 별도로 마련해야 합니다. 기후 데이터를 분석하기 위한 AI, 온라인 상 허위조작정보나 혐오표현을 탐지하는 AI 등 특정 분야에서 사용되는 비생성형 AI를 사용할 경우 그에 대한 별도의 정책과 지침이 필요합니다.

3. 생성형 AI 활용 지침

1) 정보의 정확성 확인

생성형 AI의 결과물은 부정확한 정보를 포함할 수 있으므로,
반드시 신뢰할 수 있는 방식으로 정확성을 확인해야 한다.

① 사실관계가 중요한 업무에서는 생성형 AI
활용에 특별한 주의가 필요하다
② 인터넷 검색, 전문가 자문 등 다양한 출처를
통해 사실관계를 교차검증한다.
③ 데이터나 자료가 최신 정보인지 확인한다.
④ 권위있는 출처나 공식 문서를 우선적으로 참조한다.
⑤ 최근 정보를 반영한 AI의 결과물(예 : 웹검색
기반 AI 결과물)을 우선적으로 활용한다.
⑥ 명확하고 구조화된 프롬프트를 사용하고
출처를 제공할 것을 요구한다.
⑦ 원본을 직접 읽지 않은 상태에서 요약 기능만
사용하는 것은 주의해야 한다.

생성형 AI는 원리상 자신이 학습한 데이터를 기반으로 다음에 올 단어를 확률적으로 예측하는 모델입니다. 즉, AI는 무엇이 사실인지 여부를 이해하고 답변하는 것이 아니라, 문맥상 가장 그럴듯한 문장을 생성합니다. 따라서 생성형 AI의

답변이 사실 여부를 보증하지는 않습니다. 이러한 구조적 특성 때문에 생성형 AI는 사실이 아닌 내용을 마치 진실인 것처럼 생성하는, 이른바 '할루시네이션(hallucination)' 현상을 완전히 피하기 어렵습니다. 또한, AI는 마지막으로 학습한 시점 이후에 발생한 사실이나 정보에 대해서는 알지 못합니다. 이러한 문제를 보완하기 위해 최근에는 인터넷이나 별도의 데이터베이스에서 검색을 한 후에, 이 자료를 바탕으로 답변을 생성하는 소위 RAG(Retrieval-Augmented Generation, 검색 결합 생성) 방식이 활용되기도 합니다. 그러나 학습 데이터나 인터넷 상의 콘텐츠에도 정확하지 않은 정보가 있을 수 있고, AI가 잘못된 정보를 선택할 수 있으므로 마찬가지로 주의가 필요합니다.

시민사회단체는 정확성과 신뢰성이 매우 중요합니다. 잘못된 정보나 왜곡된 사실을 전달하면 단체의 신뢰를 해칠 뿐 아니라, 관련 이슈나 캠페인에 부정적 영향을 줄 수 있습니다. 따라서 AI의 결과물을 활용할 때 '사실관계가 중요한 경우', 반드시 사실 확인 절차를 거쳐야 합니다. 예를 들어, 전반적인 문장은 그럴 듯 하더라도 법률 조항, 판례 번호, 어떤 사건의 날짜, 통계 데이터 등 세부적인 사실 관계가 잘못된 경우가 많기 때문에 반드시 확인이 필요합니다.

정확성을 확인하기 위해 다양한 방법이 활용될 수 있습니다. 앞서 언급한 것처럼 AI가 자체적으로 생성한

결과물보다 인터넷에서 검색한 최근 정보를 반영한
결과물을 활용하는 것이 바람직합니다. 물론 생성형 AI가
참조했다고 밝힌 링크가 연결되지 않거나, 예전 자료
또는 중요도가 낮은 자료를 참조할 수도 있기 때문에,
외부의 정보를 참조하여 결과물을 산출했다고 하더라도
일일이 출처의 정확성을 검증할 필요가 있습니다.
해당 주제와 관련된 공식 문서나 권위있는 출처, 학술
논문 등을 우선적으로 참조하는 것이 바람직합니다.
물론 정부, 국제기구, 공공기관에서 발간한 보고서
역시 정치적으로 편향된 관점과 왜곡된 데이터를
포함할 수 있다는 점은 인식해야 합니다.
법령이 개정되거나 특정 사건이 시간에 따라 전개될
수 있기 때문에 더 최신의 정보가 있는지 확인할
필요도 있습니다. 이러한 검증 작업에 시간과 노력이
많이 소요되기 때문에, 때로는 생성형 AI를 사용하는
것이 오히려 효과적이지 않을 수도 있습니다.

다른 생성형 AI에 유사한 질의를 해서 답변 결과를 비교하는
것도 하나의 방법입니다. 서로 다른 출처를 인용할 수 있고 만일
사실 관계가 서로 다르다면 특별히 주의해서 봐야할 것입니다.
결국 생성형 AI의 결과물에 대해 최종적인 판단을 하는
것은 단체와 이를 담당하는 활동가입니다. 제대로 판단하기
위해서는 담당자의 경험과 전문성이 필요합니다. 생성형
AI를 활용하더라도 활동가의 역량이 중요한 이유가

여기에 있습니다. 활동가가 역량을 갖추고 있지 못하다면, 인터넷 검색이나 전문가 자문 등을 통해 보완하더라도 단체가 책임질 수 없는 결과물을 내놓을 수 있습니다.

자료를 다른 언어로 번역하는 다소 기술적인 작업에서도 할루시네이션이 발생할 수 있습니다. 예를 들어 챗지피티나 제미나이 등은 과거에 비해 매우 자연스러운 번역을 제공하지만, 일부 내용을 생략하거나 번역 결과물을 마음대로 편집하기도 하며, 해당 원문에 없지만 관련된 주제의 내용을 추가하기도 합니다. 대량의 자료를 번역할 경우 이러한 오류가 커질 수 있습니다. 따라서 번역된 결과물을 반드시 원문과 대조해야 합니다. 번역의 질이나 할루시네이션의 정도는 제품이나 요금제에 따라 달라질 수 있습니다. 상용 제품의 기능은 계속 변화하기 때문에 본 가이드에서는 특정 제품에 대해 언급하려고 하지 않으며, 각 단체에서 스스로 검토해보시기를 바랍니다.

이용자가 올린 자료의 요약에서도 할루시네이션이 발생할 수 있습니다. 예를 들어, 업로드된 자료에 포함되어 있지 않지만, 유사한 주제의 다른 내용이 포함될 수 있습니다. 요약한 결과물이 정말 원 자료의 핵심을 정리했는지도 검토해봐야 합니다. 생성형 AI 요약 서비스를 지나치게 신뢰하여 원 자료를 읽지 않고 요약 서비스에만 의존하는 경우 핵심을 놓칠 위험이 있습니다. 따라서 요약 결과만 읽고 원문을

읽지 않는 것은 매우 위험합니다. 가능하면 요약문은 참고용으로만 활용하고, 중요한 문서일수록 원문을 읽을 것을 권장합니다.

명확하고 구조화된 프롬프트를 사용하면 할루시네이션을 일정하게 줄일 수 있습니다. 답변의 근거나 범위, 형식 등을 조건지워 줌으로써 AI가 마음대로 창작하는 범위를 제한하는 것입니다. 예를 들어, 다음과 같은 방법을 사용할 수 있습니다.

- 답변의 근거나 출처를 구체적으로 표시하도록 요구
- 시점의 특정 : 예를 들어, 2024년 이후의 자료만 사용
- 지리적 범위의 제한 : 예를 들어, 분석 대상을 유럽과 미국의 법제로 한정
- 모르면 모른다고 말하게 하기 : 예를 들어, 출처를 확인할 수 없는 경우는 '확인 불가'로 표시할 것.
- 출력의 형식을 구체적으로 지정 : 예를 들어, 법률을 인용할 경우 조문번호를 포함하도록 할 수 있음.

물론 이렇게 해도 할루시네이션을 완전히 방지할 수 있는 것은 아닙니다. 따라서 결과물의 정확성을 검토하는 것은 여전히 중요합니다.

2) 편향 및 고정관념(stereotype)에 대한 비판적 검토

> AI는 기존의 데이터를 학습하고 이를 모방하기 때문에, 생성형 AI의 결과물이 기존 현실의 편견, 편향, 고정관념을

그대로 반영할 수 있으므로, 이러한 결과물이
공개되거나 사용되지 않도록 주의해야 한다.

① 활동가가 생성형 AI 결과물의 편향·차별 표현을
인식할 수 있도록 정기적인 인권 교육을 시행한다.
[또는 AI 결과물 검토 담당자를 지정한다]
② 생성형 AI 활용 과정에서 문제가 될 수 있는 표현이 발견되면,
즉시 사용을 중지하고 [검토 담당자]에게 전달한다.
③ 생성형 AI에 차별적이지 않은 내용으로 다시 작성해줄
것을 요구하고, 그 결과물을 다시 검토한다.
④ 해당 생성형 AI를 서비스하는 업체에
발생한 문제에 대해 신고한다.
⑤ 차별적, 혐오적 표현을 반복적으로 생성할 경우,
해당 생성형 AI의 사용을 중단한다.
⑥ 생성형 AI에만 의존하지 않고 다른 경로를
통한 정보와 관점의 수집을 검토한다.

생성형 AI는 인터넷 상의 데이터를 대량으로 학습합니다.
이 데이터는 차별적 언어, 젠더·인종·지역 편견, 사회적
위계, 고정관점(stereotype)을 그대로 반영하고 있습니다.
예를 들어, 많은 사람들이 '미등록 이주민'이 아니라
'불법 체류자'라는 용어를 사용한다면 생성형 AI는
이를 재생산할 가능성이 큽니다. 이 과정에서 소수자나
취약계층에 대한 낙인·차별·고정관념·혐오 표현을 확대

재생산할 위험이 매우 높습니다. 이러한 결과물을 아무런 문제의식없이 활용할 경우, 공익과 인권 옹호라는 단체의 핵심 가치와 맞지 않고 단체에 대한 신뢰를 손상시키게 될 것입니다. 따라서 생성형 AI를 사용할 때 이러한 위험을 감지하고 차단하는 내부 절차가 필요합니다.

이러한 위험을 방지하기 위해서는 우선 모든 활동가가 AI 결과물에서 편향·차별을 인식할 수 있도록 정기적인 인권 교육을 시행할 필요가 있습니다. 단체에 따라 필요하다면, 특히 단체 외부로 발행되는 모든 자료를 사전에 검토하기 위한 절차를 두거나 담당자 지정을 고려할 수 있습니다. 혐오·차별 표현이 의심되는 결과물을 발견하면, 사용을 즉시 중단하고 검토 담당자에게 전달합니다. 또는 생성형 AI에 수정을 요청할 수 있을 것입니다. (예 : “이 표현은 차별적일 수 있으니, 중립적·포용적 언어로 다시 작성해줘”) 그리고 재작성된 결과물에 문제가 되는 표현이 없는지 다시 검토합니다. 심각하게 문제가 되는 표현을 생성하거나 차별적·혐오적 표현이 반복될 경우, 해당 AI 업체의 신고 또는 피드백 채널을 이용하여 문제를 제기합니다. 동일한 문제가 반복되거나 해결되지 않는다면, 해당 AI 서비스의 사용을 공식적으로 중단합니다. 이를 대체할 수 있는 도구를 검토하고, 문제 사례를 내부 기록으로 남겨 재발 방지에 활용합니다.

이와 같은 점들을 주의한다고 해도, 생성형 AI의 근본적인

한계를 인식할 필요가 있습니다. 전통적인 검색 엔진은 (검색 알고리즘이 갖고 있는 문제에도 불구하고) 여러 사이트의 목록을 제공합니다. 그러나 생성형 AI는 통상 하나의 답변을 제공하는 경우가 많으며, 이는 이용자들이 AI의 답변을 무비판적으로 수용할 위험을 높입니다. 또한, 비단 사회적 소수자와 관련된 표현에만 편향이 발생하는 것은 아니며 그 사회의 비주류적 관점, 또는 인터넷을 통해 표현되지 않은 관점이나 정보를 배제할 우려가 있습니다. 이러한 구조적 문제를 고려한다면, 생성형 AI의 답변을 비판적으로 검토하는 것으로 충분하지 않을 수 있습니다. 생성형 AI에 지나치게 의존하지 말고, 중요한 주제에 대해서는 직접 조사, 전문가 의견 등 다양한 경로로 정보와 견해를 수집할 필요가 있다는 것을 항상 염두해 두어야 합니다.

3) 개인정보 보호 및 보안

상용 생성형 AI 서비스를 이용할 경우, 프롬프트로 입력한 데이터는 AI 업체의 서버에 저장되므로 무단 접근 및 유출과 같은 보안 문제가 발생할 수 있다. 또한, 이 데이터가 모델 재학습에 사용될 경우 다른 이용자의 출력물을 통해 개인정보나 기밀정보가 유출될 가능성도 있다. 이에 정당한 법적 근거없이 개인정보가 처리되거나 단체의 기밀 정보가 유출되지 않도록 보안에 주의해야 한다.

① 프롬프트를 통해 주민등록번호, 신용카드번호, 비밀번호, 민감정보(생체정보, 성적지향 등)와

같은 개인정보를 올리지 않는다.
② 생성형 AI를 통한 개인정보의 분석이 필요할 경우 가명처리를 해야 한다.
③ 보안 등급에 따라 높은 수준의 보안이 필요한 기밀자료(예 : 피해자 인터뷰, 비공개 회의록, 회계기록 등)를 프롬프트를 통해 올리지 않는다.
④ 생성형 AI 서비스의 이용약관, 개인정보 처리방침, 보안 정책을 확인하여, 데이터 보관 기간, 프롬프트로 입력된 데이터를 AI 학습에 사용하는지 여부, 개인정보보호법 등 관련 법 준수 여부, 암호화 등 보안정책, 요금제별 보안 수준의 차이를 파악한다. 가능하면 학습 데이터 활용을 거부(옵트아웃)하는 옵션이나 해당 기능이 포함된 요금제를 선택한다.
⑤ 생성형 AI를 통해 공유한 데이터는 정기적으로 백업하고 삭제한다.
⑥ 생성형 AI가 다른 애플리케이션이나 외부의 API와 연계될 경우, 그 과정에서 불필요한 개인정보나 데이터가 전송되지 않도록, 전송되는 데이터의 범위를 확인한다.
⑦ 업무용 계정과 개인용 계정을 분리하여 사용한다.

사용자가 프롬프트에 입력한 문장, 업로드한 문서 등 데이터는 AI 업체의 서버로 전송되어 저장됩니다. 이 과정에서 여러 보안 위협이 발생할 수 있습니다. 전송 중 보안 침해가 발생할 수도 있고, AI 업체가 저장된 데이터에 무단으로 접근하거나 AI 업체의 서버가 해킹당해 데이터가 유출될 수 있습니다. 구글 드라이브와 같은 클라우드에 단체의 데이터를 저장하게

될 때 검토해야 할 보안 고려사항이 여기에도 적용됩니다.

디지털정의네트워크(구 진보네트워크센터)는 2024년에 〈2024 디지털 보안 가이드〉와 〈개인정보 안전성 확보조치 가이드〉를 발간한 바 있습니다. 시민사회단체가 준수해야 할 일반적인 보안 정책 및 개인정보에 대한 보안 정책은 이 가이드를 참고하시기 바랍니다.

생성형 AI와 관련하여 고유한 추가적 보안 위협이 있습니다. AI 업체의 서버로 전송된 데이터가 이후 AI의 재학습 과정에서 학습 데이터로 사용될 수 있습니다. 생성형 AI 기술이 학습 데이터를 그 자체로 저장하고 이를 출력에 활용하는 방식은 아니지만, 연구에 따르면 (개인정보를 포함하여) 특정 정보가 모델 파라미터에 암기되어 특정한 조건 하에 추출될 수 있습니다. 따라서 재학습된 AI의 활용 과정에서 다른 이용자의 출력에 단체의 개인정보나 기밀정보가 노출될 위험이 발생합니다.

이러한 보안 위협에 대비하기 위하여 다음과 같은 안전조치가 필요합니다.

첫째, 프롬프트를 통해 주민등록번호, 여권번호, 운전면허번호 등 개인식별번호, 신용카드번호, 비밀번호, 민감정보(생체정보, 성적지향 등)와 같은 개인정보를 올리지 않아야 합니다. 한국의 개인정보보호법은 아래와 같은 정보를 민감정보로 규정하고 있습니다. 그런데 위치정보와 같이 개인정보보호법 상 민감정보로

규정되어 있지는 않지만, 개인정보 침해가 큰 정보들이 있을 수 있습니다. 다른 나라에서 민감정보로 규정하고 있는 항목은 한국과 다를 수 있습니다. 시민사회단체라면 민감정보로 간주될 수 있는 정보들을 폭넓게 보호하는 것이 바람직할 것입니다.

> 개인정보보호법 상 민감정보(제23조) :
> 사상·신념, 노동조합·정당의 가입·탈퇴, 정치적 견해,
> 건강, 성생활 등에 관한 정보, 유전정보, 범죄경력자료,
> 생체인식정보, 인종이나 민족에 관한 정보

둘째, 앞의 원칙은 고유식별정보나 민감정보에 대한 특별한 위험성을 강조한 것일 뿐, 일반적인 개인정보를 업로드하지 않는 것이 바람직합니다. 만일 개인정보의 분석이 필요할 경우에는 가명처리를 해야 합니다. 가명처리란 이름이나 개인식별번호와 같은 일부 개인정보를 삭제하거나 암호화된 문자열로 대체하는 방법으로 원본 데이터로 되돌릴 수 있는 추가적인 정보 없이는 개인을 식별하지 못하도록 만드는 것을 말합니다.

셋째, 개인정보가 아니더라도 보안 등급에 따라 높은 수준의 보안이 필요한 데이터, 즉 유출되었을 경우 문제가 될 수 있는 기밀자료 역시 프롬프트를 통해 올리지 않도록 주의해야 합니다. 예를 들어, 피해자 인터뷰, 중요한 결정에 대한 비공개 회의록, 회계기록 등이 이에 해당할

것입니다. 보안등급을 어떻게 설정할 것인지, AI 업체를 얼마나 신뢰할 수 있는지, 단체의 위험 감내 수준 등에 대한 판단은 단체의 상황에 따라 달라질 것입니다.

넷째, 생성형 AI 서비스의 이용약관, 개인정보 처리방침, 보안 정책을 확인하여, 데이터 보관 기간, 프롬프트로 입력된 데이터를 AI 학습에 사용하는지 여부, 개인정보보호법 등 관련 법 준수 여부, 암호화 등 보안정책, 요금제별 보안 수준의 차이 등을 파악할 필요가 있습니다.
일부 해외 생성형 AI 업체의 서비스의 경우 한국의 개인정보보호법에 대한 준수가 미흡할 수 있으며, 이 경우 국내 개인정보보호법 상 보호를 제공받지 못할 수 있습니다.

요금제에 따라 개인정보 보호 수준도 달라집니다. 많은 업체들이 무료로 서비스를 제공하는 경우 (심지어 유료로 사용하고 있더라도 개인 이용자인 경우) 프롬프트를 통해 이용자가 공유한 데이터를 AI 학습에 사용하고 있습니다. 일부 업체는 이용자가 옵트아웃할 수 있는 옵션을 제공하지만 그렇지 않은 업체도 있습니다. 만일 AI 업체가 옵트아웃(즉, AI 학습 데이터로 사용하지 않는 선택지) 옵션을 제공할 경우 그것을 선택해야 합니다. 또는 보안을 위해서는 이용자가 업로드한 데이터를 AI 학습 목적으로 사용하지 않는 요금제를 선택할 수도 있습니다. 이 경우 단체에 경제적인 부담이 됩니다. 그리고 AI 서버에 저장된

데이터의 보안이 절대적으로 보장되는 것은 아닙니다.
예를 들어 2025년 11월 현재 한국에서 제공되는 주요
생성형 AI 서비스는 다음과 같은 정책을 가지고 있습니다.
OpenAI의 ChatGPT Free 나 개인 유료요금제인 ChatGPT
Plus의 경우 이용자가 입력한 데이터를 기본적으로 학습
데이터로 사용합니다. 다만, 이용자가 설정을 변경하여
옵트아웃할 수 있도록 하고 있습니다. ('설정 → 데이터
제어 → 모두를 위한 모델 개선'에서 '꺼짐'으로 설정)
반면, 기업용 요금제인 ChatGPT Team 이나 ChatGPT
Enterprise, API (개발자용) 등은 기본 설정이 옵트아웃입니다.
즉 이용자 데이터를 학습 목적으로 사용하지 않습니다.

그림 6. 챗지피티(ChatGPT) 설정 화면 : 이용자 데이터의 훈련 목적 사용 거부 옵션

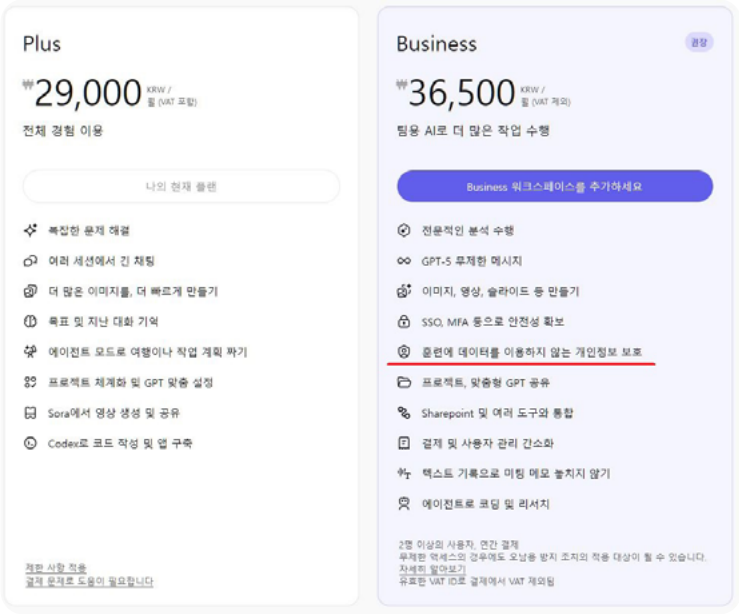

그림 7. 예시) 챗지피티(ChatGPT) 요금제 : 요금제에 따라 개인정보보호 수준이 달라짐

구글은 제미나이 독자적인 요금제가 아니라 다른 구글 서비스(구글 검색, 구글 Workspace, 구글 Cloud 등)와 통합되어 AI 서비스를 제공합니다. 챗지피티와 마찬가지로 무료 및 개인 유료 요금제의 경우에는 이용자가 업로드한 데이터를 기본적으로 AI 학습 목적으로 사용할 수 있도록 하고 있습니다. 구글 Workspace, 구글 Cloud 등 기업용 요금제의 경우 학습 데이터로 사용하지 않는다고 합니다. 구글 역시 'Gemini 앱 활동' 기능을 끔으로써 모델 학습에 사용되지 않도록 할 수 있는데, 이 경우 대화 자체를 저장하지 않게 됩니다. 즉, 챗지피티의 경우에는

대화 기록 자체를 삭제하지 않고 학습 데이터로 사용하지 않도록 하는 옵션을 선택할 수 있는데, 제미나이의 경우 옵트아웃을 선택하면 대화 기록까지 삭제하게 됩니다.

그림 8. 제미나이(Gemini) 설정 화면 : 이용자 데이터의 훈련 목적 사용 거부 옵션

클로드(Claude) 서비스를 제공하는 앤트로픽(Anthropic)의 경우 2025년 10월 8일 정책을 변경하여 이용자가 가입할 때 자신의 데이터를 AI 학습과 개선 목적으로 사용할 수 있도록 할 것인지 선택하는 옵션을 제공하고 있으며, 이후에도 설정에서 변경할 수 있습니다. 앤트로픽 역시 기업 사용자의 데이터는 AI 학습에 사용하지 않는다고 합니다.

그림 9. 클로드(Claude) 설정 화면 : 이용자 데이터의 훈련 목적 사용 거부 옵션

이처럼 생성형 AI 서비스마다, 그리고 요금제마다 프라이버시 정책이 다릅니다. 또한 이 정책은 시간이 지나면서 자주 변경이 됩니다. 따라서 단체는 자신이 사용하고자 하는 AI 서비스의 정책에 대해 세심하게 검토할 필요가 있습니다.

상용 생성형 AI 서비스를 이용하는 경우, AI 업체의 서버에 단체가 프롬프트를 통해 입력한 내역이나 업로드한 데이터가 저장된다는 점에서 보안상 취약점이 존재합니다. 이는 구글 클라우드와 같이 빅테크 업체의 클라우드 서비스를 이용하는 경우도 마찬가지입니다. 이러한 보안 위협을 피하고자 한다면 믿을 수 있는 단체/기업의 서비스를 사용하거나 자체 서버 공간에 데이터를 보관해야 합니다. 오픈소스 모델을 사용하여 독자적인 서비스를 구축하거나, 상용 생성형

AI 서비스를 이용하더라도 독자적인 시스템을 구축하는 방향으로 계약을 체결할 수 있습니다. 그러나 이를 위해서는 단체들이 이 시스템을 운영할 수 있는 기술력과 자금이 필요합니다. 안타깝지만 이를 감당할 수 있는 시민사회단체는 많지 않을 것입니다. 오픈소스 모델의 한국어 지원이 취약하다는 점도 한국 이용자에게는 하나의 장벽입니다.

개인정보 및 보안이 되는 채팅을 하고자 할 경우 Duck.ai가 하나의 대안이 될 수 있습니다. 개인정보를 보호하는 검색엔진을 표방하는 덕덕고(DuckDuckGo)는 이용자의 개인정보를 익명처리하여 앤트로픽(Anthropic)의 클로드(Clause), 메타의 라마(Llama), OpenAI의 GPT-4/5 등의 모델과 대화를 나눌 수 있는 AI 서비스인 Duck.ai를 제공합니다. Duck.ai는 사용자의 사용방식을 추적하거나 대화 내용을 저장하지 않으며(대화 내용은 원격 서버가 아닌 기기에 저장된다고 합니다) 데이터를 AI 학습에 사용하지 않는다고 합니다. IP주소와 같은 개인정보가 포함된 모든 메타데이터는 앤트로픽이나 OpenAI와 같은 모델 제공자에게 메시지를 보내기 전에 완전히 삭제됩니다. 즉, 앤트로픽이나 OpenAI는 누가 메시지를 보냈는지 알 수 없다는 것입니다. 그러나 Duck.ai를 사용하더라도 프롬프트에 포함된 개인정보나 기밀정보까지 보호되는 것은 아닙니다. 이 프롬프트의 정보는 Duck.ai를 통해 AI 업체로 전송됩니다. 다만 Duck.ai에 따르면 AI 업체에 전송된 데이터도 응답을 제공하는 데

더 이상 필요하지 않을 때(최대 30일 이내, 안전 및 법률 준수를 위한 몇 가지 예외 사항 포함) 수신한 모든 정보를 삭제한다는 내용의 계약을 체결했다고 합니다. 아직 한국에서 사용하기에는 다른 상용 생성형 AI에 비해서 기능상의 제약이 있기는 하지만, 상대적으로 좋은 보안을 제공하고 있기 때문에 용도에 따라서는 이 서비스의 사용을 고려할만 합니다.

다섯째, 생성형 AI를 통해 공유한 데이터의 보안이 우려된다면 기존에 공유한 데이터를 정기적으로 백업하고 삭제할 필요가 있습니다. 물론 삭제를 설정한다고해서 AI 업체의 서버에서 바로 삭제되는 것은 아니고, 일정기간 동안(예를 들어, 30일 정도) 남아있을 수 있습니다. 그럼에도 삭제를 하는 것이 데이터 보안 위협을 줄일 수 있습니다. 다만, 생성형 AI는 답변을 생성하는데 기존 대화내용을 참조할 수 있기 때문에, 이를 삭제할 경우 활용성을 제약할 수 있다는 점은 고려해야 합니다. 삭제 시점 및 책임자를 문서로 남겨둔다면 장기적으로 관리하는데 도움이 될 것입니다.

여섯째, 생성형 AI가 다른 애플리케이션이나 외부의 API와 연계될 경우, 생성형 AI가 필요 이상으로 자신의 데이터에 접근하거나 데이터를 외부 업체에 전송하지 않도록, 연결되는 애플리케이션이나 전송되는 데이터의 범위를 확인해야 합니다. 예를 들어 챗지피티의 GPT 탐색 기능은 여행 계획을 위해 Expedia 서비스와 이미지 작업을 위해

Canva 서비스와 연동되어 있습니다. 챗지피티 내에서 입력한 내용의 일부가 Expedia나 Canva와 같은 외부 업체에 전송되며 여기에도 개인정보가 포함될 수 있습니다. 구글 제미나이는 지메일, 캘린더, 구글 독스 등 다른 구글 서비스와 연동될 수 있으며 항공권이나 호텔 검색과 같이 외부 서비스를 이용하기도 합니다. 이때 내가 프롬프트로 입력하거나 업로드한 내용 중 어떤 부분이 외부 업체에 전송되는지 파악하기 쉽지 않습니다. 이들 AI 앱들이 스마트폰에서 작동할 경우 연락처, 위치, 사진 등 스마트폰에 저장된 내용이나 기능에 접근할 수 있습니다. 물론 각각의 허용 여부를 이용자가 통제할 수는 있지만 수많은 설정을 제대로 파악하고 관리하는 것은 쉬운 일이 아닐 것입니다.

AI가 '생성형'에서 더 나아가 내 업무를 대신하는 '에이전트'로서의 기능이 강화되면 개인정보 위협이 더욱 증가하게 될 것입니다. 내 스마트폰의 AI 에이전트와 항공사 사이트의 AI 에이전트가 데이터를 주고 받는 것과 같이 여러 개의 에이전트가 데이터를 주고받게 될 경우 정보의 흐름을 파악하는 것이 지금보다 훨씬 힘들어질 것입니다. 비록 이용자가 지시하고 중간 중간 확인한다고 해도 이용자가 지시한 작업의 수행을 위한 세부적인 단계는 에이전트가 알아서 수행하게 될 것이기 때문입니다. 이 과정에서 내 개인정보에 누가 접근하는지 전송된 개인정보가 얼마나 오래동안 보관되는지 등을 파악하기 힘들고, 제대로 관리되지

않거나 의도적인 남용에 노출될 수 있습니다. 수많은 전송 과정에서 보안이 침해될 위험성이 증가하고 자신의 계정을 에이전트가 대리 사용하게 되어 정보주체가 인지하지 못하는 사이에 계정이 조작될 위험성도 증가합니다. 이에 따라 정책적으로는 필요한 최소한의 정보만 전송되도록 한다던가 사용 목적이 다할 경우 폐기하도록 하는 등 개인정보 보호원칙의 준수가 더욱 중요해집니다. 또한 개인정보처리자인 AI 업체들이 정보주체에게 개인정보의 접근과 활용에 대해 보다 쉽게 설명해야 할 의무도 강화될 필요가 있습니다. 이는 시민사회단체가 정책 결정자에게 요구해야 할 사항이지만 당장 서비스를 사용하는 이용자의 입장에서는 이러한 위험성에 대해 인지하고 자신의 보안 정책 및 사용 패턴에 반영할 필요가 있습니다.

일곱째, 업무용 계정과 개인용 계정을 분리하는 것이 바람직합니다. 개인용 계정으로 업무를 했을 경우 향후 문제가 생겼을 경우 추적이 어려울 수 있습니다. 물론 이 경우 단체가 개인별로 계정을 제공해야 하기 때문에 단체의 비용 부담이 있을 수 있습니다.

4) 저작권

생성형 AI 사용시 여러 측면에서 저작권 침해 위험이 있다. 사회적으로는 저작권자의 데이터를 허락없이 AI 학습 데이터로 사용할 수 있는지가 논란이 되고 있지만, 이는 이용자가 통제할 수 없는 부분이다. 다만, 학습에 사용된 개인정보나 저작물이 모델에 암기되어 결과물에 반영될 수 있기 때문에, 생성형 AI의 결과물이 학습에 사용된 저작물과 상당히 유사하게 생성될 경우, 이용자의 의도와 무관하게 저작권 침해 책임을 질 수 있다.

① 생성형 AI의 결과물(특히, 이미지나 오디오)이 의도하지 않게 저작권을 침해할 수 있으므로 주의한다. 사용하기 전에 유사한 저작물이 있는지 검색(예 : 이미지 검색)해본다.
② 생성형 AI 결과물을 재료로 사람이 직접 상당한 수정·편집을 거쳐 활용한다.

생성형 AI는 학습을 위해 공개된 데이터 또는 별도의 계약을 통해 입수한 데이터 등 다양한 데이터를 사용하는데 여기에는 개인정보 뿐만 아니라 저작권이 있는 저작물도 포함될 수 있습니다. 저작물은 시·소설과 같은 어문 저작물, 음악, 사진이나 그림 등의 이미지, 영상 저작물 등 다양한 형식을 포함합니다.

저작권 보호기간이 지나 더이상 보호되지 않는 저작물도 존재합니다. 한국 저작권법에 따르면 저작재산권은 저작자 사후

70년, 업무상 저작물은 공표 후 70년간 보호됩니다. 그런데 AI 학습을 위한 저작물 이용을 둘러싼 AI 업체와 저작권자 사이의 갈등은 전 세계적으로 뜨거운 이슈이며 여러 소송이 진행 중에 있습니다. AI 업체와 저작권자 사이에서 개별 계약이 체결되기도 하지만 2025년 11월 현재 아직 명확하게 해결된 문제는 아닙니다. 이에 대해서 저작권을 보호해야 한다는 견해와 AI 학습 목적의 저작물 이용을 공정이용으로 허용해야 한다는 견해 등 다양한 의견이 존재하며, 이에 대한 상세한 논의는 본 가이드의 범위를 벗어납니다.

그러나 생성형 AI를 사용하는 과정에서 이용자인 시민사회단체가 저작권 분쟁에 휘말릴 수 있기 때문에 이에 대해서는 주의가 필요합니다. 즉, 개인정보와 마찬가지로 생성형 AI가 산출한 결과물, 특히 음악, 이미지, 영상이 AI가 학습한 타인의 저작물을 포함하여 상당히 유사해보일 수 있고 이에 대해 원 저작권자가 저작권 침해를 주장할 수 있습니다. 이 경우 생성형 AI 업체의 책임과 별개로 생성형 AI를 이용해 해당 결과물을 생성한 이용자 역시 저작권 침해 책임을 질 수 있습니다. 비록 이용자가 저작권이 있는 저작물과 유사하다는 것을 알지 못했거나 저작권 침해 의도가 없었더라도 말입니다. 따라서 단체에 대한 신뢰 훼손과 법적 분쟁을 예방하기 위하여 생성형 AI를 사용하는 이용자는 자신도 모르게 타인의 저작권을 침해하지 않도록 주의를 기울여야 합니다. 특히, 결과물을 단체 사업 홍보 등

공개적으로 사용할 경우에는 사전 점검이 필수적입니다. 이를 위해 생성형 AI의 결과물과 유사한 저작물이 있는지 확인해야 합니다. 인터넷을 검색하거나 관련 저작권 데이터베이스를 통해 검색해볼 수 있습니다. 텍스트 역시 검색엔진으로 특정 문단을 검색하여 검증할 수 있으며 이미지 역시 이미지 검색을 통해 확인할 수 있습니다.

저작권 침해 위험을 줄이는 또 하나의 방법은 생성형 AI 결과물을 활용하되 이를 사람이 직접 수정, 가공, 편집하는 것입니다. 생성형 AI가 만든 결과물은 저작권이 없지만, 사람의 창작적 기여가 더해진 경우 저작권을 인정받을 수 있다는 장점도 있습니다.

생성형 AI를 사용한 모든 내역을 기록하기는 어렵지만 저작권 분쟁이 우려되는 경우 생성형 AI 사용과 관련한 내용, 예를 들어 사용한 AI 도구명, 생성일시, 프롬프트, 수정 여부, 담당자 등의 정보를 기록해두면 향후에 대응하는데 도움이 될 수 있습니다.

5) 생성형 AI 활용의 투명성

> 생성형 AI를 사용하는 사실을 수용자들이 인지하지 못해 오해나 혼란을 야기할 우려가 있을 경우, 해당 결과물이 생성형 AI를 통해 만들어졌음을 표시한다.

① 생성형 AI를 활용한 분석, 또는 생성형 AI로 제작한 음악, 이미지, 영상 등 생성형 AI가 결과물 생성에 핵심적으로 기여한 경우, 해당 저작물이 생성형 AI에 의해 만들어졌다는 사실을 표시한다.
② 딥페이크와 같이 생성형 AI를 이용해 현실과 혼동될 수 있는 결과물을 생성한 경우, 그 사실을 저작물에 표시한다. 다만, 예술적·창작적 저작물의 경우, 감상을 저해하지 않는 방식으로 표시할 수 있다.
③ 챗봇, 동시통역 도구 등 외부 사람과 직접적으로 상호작용하는 생성형 AI의 경우, 사람들이 자신과 상호작용하고 있는 대상이 AI라는 점을 명확하게 인지할 수 있도록 알린다.
④ 홈페이지 등을 통해 본 단체의 생성형 AI 정책을 공개한다.

앞서 원칙 부분에서 설명했듯이, 생성형 AI 활용 과정에서도 책임성과 투명성 원칙은 매우 중요합니다. 그러나 생성형 AI를 조금이라도 활용한 모든 결과물에 일일이 그 사실을 표시하는 것은 현실성이 없고 수용자에게도 큰 의미가 없습니다. 물론 이는 생성형 AI의 결과물이 단체의 책임 하에 엄격하게 검토되었다는 것을 전제로 합니다. 만일 생성형 AI로 보고서를 만든 후에 사실 확인이나 편향 여부를 전혀 검토하지 않은 채 외부에 공개했을 경우 해당 보고서는 사실 관계가 틀렸거나 편향된 내용을 포함할 수 있습니다. 그럼에도 생성형 AI 사실 사실까지 표시하지 않는다면 수용자는 보고서의 모든 내용을 사실로 받아들일 수 있고

향후에 잘못된 내용이 드러날 경우 단체의 신뢰성을 심각하게 저해할 수 있습니다. 반대로 잘못된 내용이 드러나지 않으면 오히려 허위나 편향적인 정보가 더 확산될 것이고, 단체 그 책임에서 자유로울 수 없습니다. 따라서 단체는 책임성을 갖고 정보의 정확성과 편향 가능성을 확인하기 위해 최대한 노력해야 하며 이를 보장하기 힘들 경우 최소한 해당 결과물이 생성형 AI를 활용해 작성되었고 일부 내용에 오류가 있을 수 있다는 점에 대해 수용자에게 알리는 것이 바람직합니다.

이러한 고지는 단체에서 내용에 대한 검토를 일정하게 수행한 경우에도 적용될 수 있습니다. 예를 들어, 생성형 AI를 활용해 데이터를 분석한 경우 사람이 모든 오류 가능성을 찾아내기 어렵습니다. 예술적, 창작적 결과물인 경우 아무런 표시가 없다면 인간이 만든 작품으로 오해될 수 있습니다.

아직 생성형 AI의 결과물이라는 것이 어느 정도 눈에 띄는 경우가 많지만 기술이 고도화될수록 그 경계는 희미해질 것입니다. 심지어 딥페이크와 같이 현실과 혼동할 수 있도록 의도적으로 조작된 이미지나 영상이라면 수용자의 혼란을 넘어 추가적인 문제를 야기할 수 있습니다.
딥페이크 기술은 딥페이크 성폭력물처럼 불법적 용도뿐 아니라 합법적인 저작물의 제작에 활용될 수도 있습니다. 예를 들어, 시민사회단체에서 성소수자의 신원을 보호하기 위해서 다큐멘터리에서 활용하거나 권력자를 비판하는

패러디물을 제작하는데 활용할 수도 있습니다. 이 경우 표시 의무가 작품의 감상을 방해한다면 작품의 향유를 방해하지 않는 방식(예를 들어 크레딧에 표시)으로 표시할 수도 있습니다. 딥페이크에 그 사실을 표시하는 것은 EU AI Act 준수를 위한 요건이기도 하고 점점 더 많은 나라에서 유사한 규제를 도입할 가능성이 높습니다.

6) AI가 환경에 미치는 영향에 대한 고려

생성형 AI의 확대에 따라 데이터센터 운영을 위한 전력사용량과 물 사용의 증가, AI를 위한 반도체 생산 등 자원사용량이 증가하고 있다. 이에 환경에 미치는 부정적 영향을 최소화할 수 있는 방식으로 AI를 활용한다.

① 감사인사 등 불필요한 대화 또는 에너지를 많이 사용하는 이미지·음성·영상 처리 요청을 지양한다.
② 동일한 자료를 재요청하는 경우가 많을 때는 생성된 결과물의 재활용, 단체 구성원간 결과물 공유 등을 통해 불필요한 반복 요청을 최소화한다.
③ 생성형 AI가 아니어도 처리가 가능한 업무는 적합한 다른 도구를 우선 활용한다.
④ 가능하다면 경량 AI 모델을 사용한다.
⑤ AI 운영을 위한 데이터센터의 환경영향평가, 전력사용량, 에너지효율 등의 정보공개, 친환경 재생에너지 사용 등 친환경 정책을 실천하는 기업의 제품을 사용한다.

앞서 설명했듯이, 생성형 AI의 개발 및 운영에는 막대한 연산 자원과 에너지가 필요합니다. 따라서 생성형 AI를 사용할 때 환경적 영향을 고려하는 것도 매우 중요한 인권적 실천입니다. 시민사회단체들은 AI의 개발 및 운영 과정에서 사용되는 에너지 데이터를 공개할 것, 친환경 에너지를 사용할 것, 무분별한 데이터센터 설립을 자제할 것 등 AI가 환경에 미치는 부정적 영향을 최소화하기 위해 다양한 요구를 제기해왔습니다. 하지만, AI 제공업체가 환경에 미치는 영향에 대해 '이용자 입장'에서 개입하는 것은 쉽지 않을 수 있습니다. 그럼에도 불구하고 우리가 실천 가능한 방법들을 계속 모색하는 것은 여전히 중요합니다.

우선 불필요한 생성형 AI 활용을 줄이는 노력이 필요합니다. 물론 어느 정도가 필요 최소한의 사용인지 명확한 것은 아니지만 AI 활용시 항상 환경적 영향을 염두에 두어야 하겠습니다. 예컨대 챗봇에 감사 인사를 하는 등 불필요한 대화를 하지 않도록 하고 특히 텍스트보다 훨씬 많은 전력을 소비하는 이미지·음성·영상 생성은 꼭 필요한 경우가 아니라면 지양하는 것이 좋습니다. 한 단체 내에서 동일한 요구를 반복적으로 발생한다면 기존에 생성한 결과물을 재활용하거나 구성원간 공유를 통해 불필요한 요청을 줄일 수 있습니다. 물론 저장된 자료의 최신성과 정확성을 검토하고 단체 내 공유 과정에서 부서 간 부적절한 개인정보 공유가 발생하지 않도록 주의해야 합니다. 일반적인 검색

엔진을 이용하거나 데이터 분석을 위한 오프라인 도구를 이용하는 등 생성형 AI가 아니어도 처리가 가능한 업무는 적합한 다른 도구를 우선 활용할 필요가 있습니다.

가능하다면 경량 AI 모델(예를 들어, ChatGPT 4o 대신 4o mini, Claude Opus 대신 Haiku)을 사용할 수 있습니다. 경량 모델은 대규모 모델보다 전력 소비와 연산 부하가 훨씬 적습니다. 단순 요약, 정리, 번역, 분류 등 굳이 초대형 모델이 필요하지 않은 경우가 많습니다. 환경 부담을 줄이기 위해 적절한 용도에 맞는 모델을 선택할 필요가 있습니다. 물론 어떤 모델이 적절한지 이용자가 판단하기 어려울 수 있으며 AI 업체들이 이를 자동적으로 판단할 수 있는 인터페이스를 개발하는 것이 효과적일 수 있습니다.

또한 AI 운영을 위한 데이터센터의 환경영향평가, 전력사용량, 에너지효율 등의 정보공개, 친환경 재생에너지 사용 등 친환경 정책을 실천하는 기업의 제품을 사용할 필요가 있습니다. 어떤 기업이 친환경 정책을 실천하고 있는지 판단하기 위해서는 우선 기업들이 관련 데이터를 투명하게 공개할 필요가 있습니다. 기업의 환경정책이나 ESG 보고서 등 기업의 데이터를 참조할 수 있을 것입니다.

4. 정책의 수립과 집행

1) 생성형 AI 사용 승인

① 단체의 사업 목적으로 생성형 AI를 사용하기 위해서는 [운영위원회]의 승인을 거쳐야 한다.
② 특정 생성형 AI 사용을 승인하기 전에, 해당 AI 서비스의 성능, 적절한 요금제, 필요한 설정 등 사용 정책을 마련한다.
③ AI 책임자는 본 단체에서 활용하는 생성형 AI의 목록을 관리하고, 변경시 구성원에게 공지한다.
④ 생성형 AI로 인해 업무를 대체하거나 변경할 필요가 있을 경우, 단체의 구성원과 사전에 협의한다.

단체 구성원 개개인이 임의로 다양한 AI 서비스를 사용할 경우, 신뢰할 수 없는 AI 서비스를 사용하게 되거나 본 정책과 일치하지 않는 방식으로 이용할 위험이 있습니다. 이러한 위험을 단체 차원에서 체계적으로 통제하기 위해서는 단체에서 사용할 AI 도구를 승인하고 관리하는 절차가 필요합니다.

이를 위해 특정 AI 서비스를 사용하기 전에 해당 AI 서비스의 성능에 대해 자세히 파악할 필요가 있습니다. 요금제에 따라 달라지는 사항을 검토하고 본 정책을 준수하기 위해 필요한 설정은 무엇인지, 사용하지 말아야 하는 기능은 무엇인지 등을 포함한 사용 정책을 마련해야 합니다.

운영위원회 등 적절한 단체 내 의사결정기구에서 특정 생성형 AI 사용 여부를 결정하고 단체 AI 책임자가 이 목록을 관리합니다. 이 목록에는 해당 AI 서비스 이름, 제공업체, 버전, 요금제, 사용정책, 승인일 등을 기록할 수 있습니다.

생성형 AI가 내부 구성원이 수행하던 기존의 업무를 일정하게 대체하거나 변경할 가능성이 있는 경우, 당사자들과 사전에 협의해야 할 필요가 있습니다. 노동과 인권의 가치를 중요하게 생각하는 시민사회단체라면, 생성형 AI를 도입할 때도 이를 고려하는 방식으로 접근해야 합니다. 생성형 AI로 특정 업무를 수행하던 활동가의 노동을 일방적으로 대체하는 것이 아니라, 생성형 AI가 어떠한 변화를 가져올지, 그에 따라 사람의 역할을 어떻게 재설계할지, 그에 따른 부담이나 이익을 어떻게 분배할지에 대해 협의해야 합니다. AI를 활용해 업무를 효율화한다고 판단했지만 AI가 대체할 수 없는, 생각하지 못한 다른 문제가 있을 수도 있습니다. 단순 반복 업무를 AI를 통해 효율화하되 그에 따른 새로운 역할을 부여할 수도 있고 사람과의 관계 형성 등 인간에게만 가능한 업무로 재구성할 수도 있습니다.

2) 생성형 AI 활용이 허용되는 활용의 범위

> AI 책임자는 본 단체에서 생성형 AI의 활용이 허용되는 사례, 허용되지 않는 사례, 또는 엄격한 검토가 필요한 활용 사례를 문서로 관리한다.

단체의 정책적 입장이 강하게 드러나는 업무, 개인정보 및 보안이 중요한 업무 등을 생성형 AI에 의존하는 것은 부담스러울 수 있습니다. 이러한 업무에 활용하는 것은 사전에 제한하거나 또는 엄격한 검토를 거치도록 할 필요가 있습니다. 이와 같이 허용, 제한, 엄격한 검토가 필요한 활용의 범위를 사전에 규정해두면 단체 구성원이 일관된 원칙을 가지고 생성형 AI를 활용할 수 있을 것입니다. 물론 단체의 활동 방식이나 가치에 따라 어떤 업무에 생성형 AI를 사용할지 달라질 수 있습니다. 예를 들어, 어떤 단체는 단체의 메시지가 드러나는 성명서 작성을 생성형 AI에 의존하는 것은 옳지 않다고 판단할 수 있습니다. 반면 기존에 유사한 이슈에 대해 단체가 성명을 자주 발표해왔고 최종적인 검토를 사람이 한다면 생성형 AI의 도움을 받을 수 있다고 판단할 수도 있습니다.

각 단체는 자유로운 형식으로 작성할 수 있지만 예를 들어 다음과 같이 허용되는 활용, 엄격한 검토가 필요한 활용, 허용되지 않는 활용의 목록을 문서로 관리하여 단체 구성원이 공유하도록 합니다.

아래의 내용은 이 가이드에서 권장하는 것이 아니며, 단순한 예시일 뿐입니다.

허용되는 활용	엄격한 검토가 필요한 활용	허용되지 않는 활용
자료 번역 회의록 녹취 및 요약 자료 검색 아이디어 구상	연구 보고서 작성 캠페인 홍보물 작성 법률 자문 및 분석	성명서 및 컬럼의 작성 이미지와 영상 제작 피해자 인터뷰 분석 회원 개인정보 분석

3) 교육 및 역량 강화

> ① 모든 구성원이 본 정책을 숙지하고, AI 관련 최신 동향에 대해 인지할 수 있도록 [1년에 1회 이상] 구성원에 대한 AI 교육을 시행한다.
> ② 업무상 필요한 도구의 사용법에 대한 교육의 일환으로, 생성형 AI의 사용법에 대한 교육을 시행한다.
> ③ 단체 구성원의 역량 강화를 위해 필요할 경우, 업무 수행 과정에서 생성형 AI의 도움을 받는 것을 제한할 수 있다.

단체의 AI 정책이 실제로 작동하기 위해서는 단체 구성원이 이를 이해하고 이행해야 합니다. 본 정책을 수립하는 과정에서부터 구성원이 함께 토론할 필요가 있으며 새로 합류하는 구성원을 고려하면 정기적인 교육이 필수적입니다. 이때 정책에 대한 이해와 변경 필요성에 대한 토론을 원활하게 하기 위해, AI 관련 최신 동향도 함께 교육을 하는 것이 도움이 될 것입니다. 할루시네이션, 편향 등 생성형 AI 결과물의 근본적인 문제를 이해하기 위해서는 AI의 기술적 특성에 대한 교육도 어느정도 필요할 수 있습니다. 단체 내외부에서 발생한 문제 사례(예를 들어, 편향적 결과를 도출한 사례)를 축적하고 이를 공유하는 것도 문제를 체감하는데 도움이 될 것입니다. 본 가이드가 단체 내 교육에 좋은 참고 자료가 될 수 있기를 바랍니다.

단체 차원에서 생성형 AI를 활용하기로 한다면 AI 활용 능력에 있어서 활동가 사이의 격차가 발생하는 것은 바람직하지 않을 것입니다. 이 때문에 생성형 AI의 사용법에 대한 교육이 필요할 수 있습니다. 생성형 AI를 특별하게 취급할 필요는 없고, 단체 내에서 업무상 필요한 도구의 사용법에 대한 통상적인 교육의 일환으로 시행하면 될 것입니다.

때로는 단체 구성원이 특정한 생성형 AI를 사용하는 것을 일정기간 동안 정책적으로 제한할 수도 있습니다. 생성형 AI의 결과물의 편향이나 오류를 제대로 판단하고 검토하기 위해서는 그에 합당한 전문성과 경험이 필요합니다. 따라서 그런 역량을 갖추고 있지 못한 활동가에게 생성형 AI를 활용하도록 하는 것은 적절하지 않을 수 있습니다. 또한 단체에 따라 신입 활동가의 역량 강화를 위해 성명서를 직접 작성하거나 회의록을 정리하는 업무를 맡기기도 합니다. 이를 생성형 AI가 대신한다면 신입 활동가의 학습과 역량 강화에 아무런 도움이 되지 않을 것입니다. 따라서 단체 차원에서 생성형 AI의 활용 자체를 전면적으로 제한하지는 않더라도, 특정한 구성원을 대상으로 일정 기간 동안 업무에 생성형 AI를 사용하는 것을 제한하는 정책을 가질 수 있습니다.

4) 외부 파트너와의 협력

> 다른 단체 또는 외부 사람들과 협업을 하거나, 기고를 받는 등 단체의 활동을 위해 협력할 때, 생성형 AI 활용 정책에 대해 외부 사람에게 사전에 고지하거나, 해당 정책에 대해 협의해야 한다.

시민사회단체는 다른 단체와 연대 활동을 하거나 외부의 필자, 프리랜서, 전문가 등과 협업하는 경우가 많습니다. 단체 내부의 생성형 AI 정책이 다른 단체 및 외부 파트너와 합의되지 않을 경우, 공동의 결과물이 단체의 정책에 위배되거나 단체의 신뢰에 피해를 입히는 상황이 발생할 수 있습니다. 예를 들어, 외부 필자가 생성형 AI로 작성한 원고에 정확하지 않은 내용들이 포함될 수 있습니다. 해당 결과물이 단체의 이름으로 공개된다면 단체 역시 그 책임을 피하기 어렵습니다. 따라서 사전에 단체의 생성형 AI 정책을 공유하고 동의를 받거나, 정책에 대한 이견이 있을 경우 협의하는 과정이 필요합니다. 원고나 디자인 등 특정한 업무나 결과물을 의뢰할 경우, 요청서나 협약서에 “본 단체의 AI 활용 정책을 준수할 것”이라는 문구를 포함할 수 있습니다.

5) 문제발생 시 조치

① 생성형 AI와 관련되어 문제가 발생할 경우 즉시 'AI 책임자'에게 보고한다. 보고에는 다음과 같은 내용을 포함한다 : 발생 일시 / 사용 도구명 / 해당 결과물 / 문제가 된 부분 / 프롬프트 입력 내용 / 부정적 영향의 내용과 범위
② AI 책임자는 즉시 사실을 확인하고, 필요할 경우 피해 확산을 방지하기 위한 긴급조치를 취한다.
③ AI 책임자는 [운영위원회]를 소집하여 단체의 대응 방안을 수립한다. 이를 위해 문제의 원인, 영향의 범위, 단체의 책임 유무 및 범위, 관련 법제 및 법적 대응의 필요성 등을 검토한다.
④ 필요할 경우 적절한 방식으로 해당 사안에 대해 외부에 공지한다. 공지에는 문제의 내용 및 원인, 영향을 받는 당사자, 단체의 대응 조치, 재발 방지 조치 등이 포함될 수 있다.
⑤ 필요할 경우 적절한 방식으로 영향을 받는 당사자에게 사과문을 전달한다. 사과문에는 문제의 내용 및 원인, 단체의 대응 조치, 피해 구제 및 보상, 재발 방지 조치 등이 포함될 수 있다.
⑥ 재발 방지 대책을 수립하고 필요하다면 본 정책에 반영한다
⑦ AI 책임자는 본 사안에 관련된 모든 내용과 과정을 기록한다.

본 정책의 첫번째 원칙에서 밝힌 바와 같이, 생성형 AI의 결과물로 인한 모든 책임은 우리 단체에 있습니다. 문제가 발생할 경우 단체에 대한 신뢰가 훼손되고 피해자가 발생할 수 있는데, 이에 대해 적절하게 대응하지 못하면 단체에 대한 신뢰는 더욱 악화될 것입니다. 생성형 AI로 인한 문제가 발생할

경우의 대응 절차를 마련해놓지 않으면, 어떻게 대응해야할지 우왕좌왕하거나 임시방편적 대응을 할 위험이 있습니다.

기본적으로 생성형 AI로 인해 발생한 문제에 대한 대응 절차는 다른 원인으로 발생한 문제에 대한 대응 절차와 크게 다르지 않습니다. 문제가 발생하면 책임자에게 보고하고, 즉시 사실 확인에 들어가야 합니다. 보안 문제와 같이 즉각적인 대응이 필요할 경우에는 원인에 대한 사실 확인이 늦어지더라도 피해 확산을 막기 위한 긴급조치를 취해야할 수 있습니다. 책임있게 문제를 해결할 수 있는 단위(예를 들어 운영위원회)에 보고하고 구체적인 대응 방안을 수립해야 합니다. 필요할 경우 해당 사안을 외부에 공지하고 사과를 해야할 수도 있습니다. 저작권 침해와 같은 특정한 피해자가 있을 경우 당사자에게 사과하고 적절한 보상을 제공해야 할 것입니다. 어느 정도 문제가 수습된 이후에는 재발을 방지하기 위해 정책에 반영해야 할 사항이 있는지 검토합니다. 그리고 이 모든 과정과 관련 자료를 기록으로 남겨야 합니다.

이러한 기본적인 대응 절차를 바탕으로 생성형 AI를 고려하여 세부적인 대응 절차를 마련해야 할 것입니다. 예를 들어, AI 책임자가 사고에 대한 기본적인 대응을 책임지도록 할 수 있을 것입니다. 그리고 사고 발생시 보고해야 할 내용에 발생 일시, 사용 도구명, 해당 결과물, 문제가 된 부분, 프롬프트 입력 내용, 부정적 영향의 내용과 범위 등을 포함하도록 할 수 있습니다.

6) AI 책임자와 감독

> ① 본 단체의 책임있는 AI의 활용과 감독을 위해
> 'AI 책임자'를 둔다. 본 단체의 AI 책임자는 [ ]이다.
> ② 생성형 AI의 결과물이 본 단체의 정책에 부합하지 않거나,
> 본 정책을 위반하는 사례가 있을 경우 AI 책임자에게 보고한다.
> ③ 단체 구성원이 본 정책을 위반할 경우 내부 징계 절차에 따른다.

개인정보보호법에 따라 단체가 '개인정보보호책임자'를 두는 것과 마찬가지로, 인공지능 관련 정책의 수립과 집행, 감독을 책임지는 AI 책임자를 둘 수 있습니다. AI 책임자가 다른 직책을 겸임할지, 또는 AI를 담당하는 별도의 팀을 둘 것인지 등은 각 단체의 규모 및 AI를 사용하는 정도와 맥락에 따라 달라질 수 있습니다. AI 책임자는 단체의 AI 정책 수립 과정을 총괄하며, 문제가 생겼을 때 이에 대한 대응을 책임집니다. 따라서 생성형 AI의 결과물이 본 단체의 정책에 부합하지 않거나 본 정책을 위반하는 사례가 있을 경우 AI 책임자에게 보고되어야 합니다. 단체 구성원이 본 정책을 위반하여 징계가 필요할 경우에는 단체의 내부 징계 절차에 따르면 되기 때문에 본 정책안에서 별도로 다루지는 않았습니다.

7) 정책의 변경

① AI 기술의 급속한 발전을 고려하여 본 정책은 AI 책임자가 필요하다고 판단할 경우, 또는 최소한 연 1회 재검토 및 업데이트 되어야 한다.
② AI가 단체에 미치는 영향에 대해 정기적으로 평가한다.
③ 본 정책에 대해 논의할 때 모든 구성원이 참여할 수 있도록 한다.

AI 기술의 급속한 발전과 다양한 서비스의 등장을 고려할 때, AI 정책은 자주 업데이트될 필요가 있습니다. 당분간은 최소한 연 1회 재검토하도록 하되, AI 책임자가 필요하다고 판단할 경우 언제든지 재검토되어야 합니다. 특히, 생성형 AI의 결과물로 인한 사고가 발생할 경우 정책에 문제점이나 공백은 없었는지 검토할 필요가 있습니다. 이러한 재검토 과정이 없다면 정책이 기술 변화에 뒤처져 실효성을 잃거나 단체의 활동에 과도한 부담을 줄 수 있습니다. 정책을 재검토할 때에는 본 정책이 단체에 미치는 영향, 즉 단체의 구성원과 활동에 어떠한 영향을 미치는 지도 평가되어야 합니다. 이 과정에서 구성원들이 제대로 준수할 수 없는 과도한 규정은 없는지도 점검할 필요가 있습니다. 처음 정책을 수립할 때부터 정책을 재검토하는 과정의 전 주기에 단체의 모든 구성원이 참여하여 함께 논의할 수 있도록 합시다. 그래야 정책의 문제의식에 대한 구성원 간 인식을 일치시킬 수 있고 변경된 내용들이 공유되지 않아 발생할 수 있는 혼란을 예방할 수 있습니다.

참고자료

한국

- 정보인권연구소.「인공지능에 대한 인권기반접근 : 영향 받는 사람들의 인공지능」. https://idr.jinbo.net/2762 (2025)
- 정보인권연구소.「주요 분야 인공지능 정책 및 이슈 연구 : 공공, 법집행, 교육, 사회복지 분야」. https://idr.jinbo.net/2294 (2025)
- 국가정보원 생성형 AI 보안 가이드 : https://www.ncsc.go.kr:4018/main/cop/bbs/selectBoardArticle.do?bbsId=InstructionGuide_main&nttId=54340&pageIndex=1
- 행정안전부, 챗GPT 활용방법 및 주의사항 안내서 : https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=100278
- 인사혁신처 AI 활용 가이드 : https://www.data.go.kr/data/15142458/fileData.do?recommendDataYn=Y

해외

- Amba Kak and Sarah Myers West, “AI Now 2023 Landscape: Confronting Tech Power,”AI Now Institute, https://www.ai-now.local/2023-landscape (2023)
- Aiha Nguyen and Alexandra Mateescu, Generative AI and Labor: Power, Hype, and Value at Work, Data & Society, https://doi.org/10.69985/gksj7804 (2024)
- EPIC. Generating Harms. https://epic.org/generating-harms/ (2023, 2024)
- OECD Artificial Intelligence Public Observatory. https://oecd.ai/en/.

- Artificial intelligence tools: a guide for CSOs : 우크라이나 시민사회단체에서 만든 시민사회를 위한 AI 가이드 https://cedem.org.ua/en/library/ai-guide-csos/
- City of Boston Interim Guidelines for Using Generative AI : 보스턴시 생성형 AI 가이드라인 https://www.boston.gov/sites/default/files/file/2023/05/Guidelines-for-Using-Generative-AI-2023.pdf
- CyberPeace Institute Approach to Responsible Use of Artificial Intelligence : https://rai-toolkit.github.io/readings/report/CyberPeace-Institute-Approach-to-Respons/
- When AI Gets It Wrong: Addressing AI Hallucinations and Bias : https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/
- Artificial Intelligence(AI) for Nonprofits - Best Practices : https://perlmanandperlman.com/artificial-intelligenceai-for-nonprofits-best-practices
- Civil Tech Field Guide - Civil AI : https://civictech.guide/ai/
- People Powered AI Policy 2025 : https://app.civictech.guide/p/people-powered-ai-policy-2025/r/recJfYx6zp9lshdua
- Artificial Intelligence (AI) Adoption by Civil Society Organizations (CSOs) in Zambia - A Survey Report : https://internews.org/wp-content/uploads/2024/12/AI-CSO-Survey-report-validation-with-changes-proofread-03.pdf
- Grassroots and non-profit perspectives on generative AI : https://www.jrf.org.uk/ai-for-public-good/grassroots-and-non-profit-perspectives-on-generative-ai

시민사회를 위한

생성형 AI 가이드