Business Software Alliance

콘텐츠 진위성에 관한 개요

인공지능 생성 콘텐츠에 대한 투명성은 신뢰할 수 있는 인공지능을 실현하는 데 있어 핵심적인 요소입니다.

Business Software Alliance (BSA)는 이용자가 인공지능 생성 콘텐츠의 이력과 출처를 손쉽게 확인할 수 있는 신뢰성 있는 콘텐츠 인증·출처 추적 체계의 개발과 도입을 지지합니다. 이러한 체계를 통해 소비자는 특정 콘텐츠가 사람 혹은 인공지능이 만든 것인지를 보다 쉽게 파악할 수 있으며, 오정보·허위정보 문제에 대응하는 데 도움이 될 수 있습니다.

오늘날 전 세계 정책 입안자들은 공통된 과제에 직면해 있습니다. 소비자가 온라인에서 마주하는 사진과 영상이 진짜인지를 어떻게 판단할 수 있느냐는 문제입니다. 이를 위해서는 소비자가 워터마크, 디지털 지문, 보안 메타데이터처럼 이미 존재하는 도구들을 잘 인식하고 활용할 수 있어야 합니다. 이와 같은 도구들은 콘텐츠를 누가 만들었는지, 인공지능과 같은 디지털 기술로 수정이 이루어졌는지를 확인하는 데 유용합니다. 이에, 정책 입안자들은 소비자가 사람이 만든 콘텐츠와 인공지능이 만든 콘텐츠를 구별할 수 있도록 도구들의 활용을 적극적으로 뒷받침해야 합니다.

BSA의 콘텐츠 진위성에 관한 개요서는 다음과 같은 내용을 담고 있습니다:

» 인공지능 투명성 정책의 토대가 되는 콘텐츠 진위성 및 출처 개념을 정의합니다.

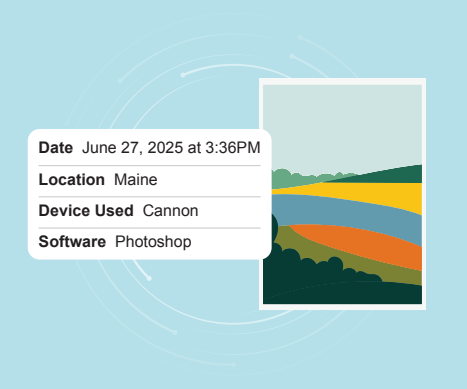
» 이미지·영상의 제작자 확인 및 변조 여부 식별에 활용할 수 있는 기존 도구들을 소개합니다. 기계 판독형 워터마크, 디지털 지문, 보안 메타데이터 등이 이에 해당합니다.

» 콘텐츠 진위성·출처 추적 도구의 활용을 뒷받침하기 위해 각 유형의 사업자가 취할 수 있는 구체적인 조치를 설명합니다.

» 딥페이크로 인해 제기되는 정책적 쟁점을 검토합니다.

» 소비자가 온라인에서 사실과 허위를 가려낼 수 있도록 다양한 인공지능 정책이 어떤 역할을 할 수 있는지 설명합니다.

콘텐츠 출처 및 인증이란 무엇인가?



콘텐츠 출처 (Content Provenance)

콘텐츠 출처 정보는 콘텐츠가 어디서 비롯되었고 어떠한 과정을 거쳐 변화해 왔는지 나타내며, 이미지, 영상, 오디오 클립 등 디지털 파일의 최초 생성 경위, 소유권 이전 내역, 수정 이력 등을 추적합니다. 예시: 사진은 촬영 일시 및 장소, 사용된 카메라 기종, 파일 편집에 사용된 소프트웨어 등이 표시된 암호화 서명된 메타데이터 정보가 포함되어 있습니다.

콘텐츠 인증 (Content Authentication)

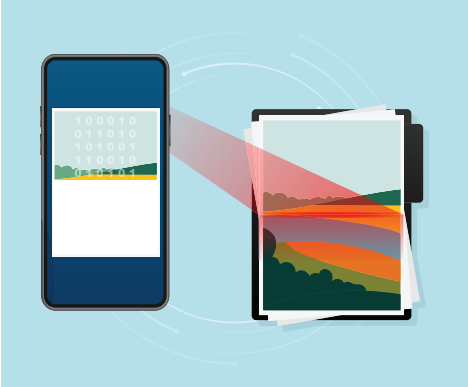
콘텐츠 인증은 해당 콘텐츠와 그 출처 정보가 신뢰할 수 있는 것인지, 메타데이터·워터마크·자격증명 등이 위·변조되지 않았는지를 확인하는 데 활용됩니다. 예시: 콘텐츠 인증 도구를 통해 영상에 삽입된 서명이 발급자의 공개키와 일치하는지, 그리고 해당 파일이 게시 이후 변경되지 않았는지를 검증할 수 있습니다.

AI 생성 콘텐츠를 식별하는 데 활용할 수 있는 도구는 무엇인가?

인공지능으로 생성된 콘텐츠를 포함하여, 콘텐츠의 생성 방식이나 수정 여부를 파악할 수 있는 다양한 기술이 이미 존재합니다. 정책 입안자들은 특정 기술의 사용을 강제하기보다는, 사업자들이 콘텐츠의 인공지능 생성 여부를 표시하는 데 여러 기술을 자유롭게 활용할 수 있는 환경을 마련하는 데 초점을 맞춰야 합니다.

기계 판독형 워터마크

기계 판독형 워터마크는 소량의 정보를 파일 안에 직접 심는 방식으로, 전용 도구를 사용하여 이를 탐지할 수 있습니다. 워터마크는 파일의 픽셀이나 데이터 스트림에 내장되기 때문에, 나중에 메타데이터가 삭제되더라도 파일의 출처를 추적·검증할 수 있습니다. 예시: 이미지에 삽입된 비가시성 워터마크를 활용하면 해당 이미지의 출처를 특정 제작자나 진위성 플랫폼으로 소급할 수 있습니다.

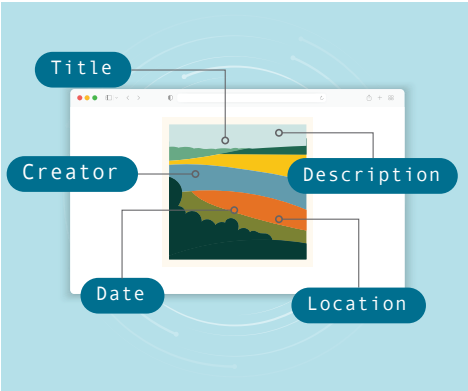


디지털 지문

디지털 지문은 색상 패턴, 구조, 인코딩 방식 등 콘텐츠의 핵심 특성을 분석해 고유 식별자, 즉 '지문'을 만들어냅니다. 디지털 지문은 파일 자체를 건드리지 않으며, 해시처럼 파일과 별도로 저장할 수 있는 고유한 수학적 서명입니다. 파일의 디지털 지문을 데이터베이스에 저장된 지문과 대조하면 해당 파일이 원본 그대로인지, 아니면 변조되었는지를 확인할 수 있습니다. 예시: 영상을 프레임 단위로 나누어 영상·음성 정보를 기반으로 지문을 생성한 뒤, 이를 디지털 서명 데이터베이스와 대조하면 해당 영상의 수정·변조 여부를 확인할 수 있습니다.

보안 메타데이터

보안 메타데이터는 파일을 누가 만들었는지, 어떤 도구를 사용했는지, 어떻게 편집되었는지에 관한 정보를 담는 데 활용됩니다. 이러한 정보는 곧 해당 파일의 콘텐츠 출처를 뒷받침하는 근거가 됩니다. 예시: 이미지에 제작자 정보, 편집 이력, 사용 소프트웨어 등을 기록한 메타데이터를 포함시키고, 이를 디지털 서명하여 파일에 내장함으로써 열람자가 해당 콘텐츠의 진위성과 출처를 검증할 수 있습니다.

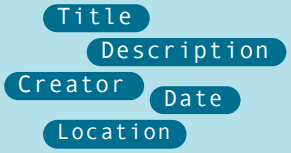


종합적 접근: 다양한 도구의 병행 활용

이러한 기술들은 서로 결합함으로써 콘텐츠 진위성 확보를 위한 견고한 체계를 만들어낼 수 있으며, 콘텐츠가 어떻게 만들어졌고 변조는 없었는지를 이용자 스스로 판단할 수 있는 실질적인 수단이 됩니다. 세 가지 방식은 각각 하나의 검증 층위로서 기능합니다.

워터마킹
진위성을 파일에 내장

디지털 지문
콘텐츠를 식별

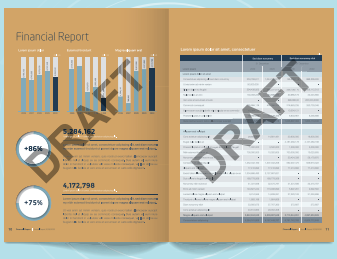


메타데이터
맥락과 저작자 정보를 기록

세 가지를 모두 적용하는 것이 바람직한 경우도 많지만, 모든 상황에서 세 가지를 빠짐없이 사용해야 하는 것은 아닙니다. 정책은 현시점에서 충분히 성숙하여 실제 도입이 가능한 기술적 방법의 활용을 장려하면서도, 다양한 기술적·운영적 환경을 폭넓게 수용할 수 있도록 유연하게 설계되어야 합니다.

AI에 맞지 않는 방식: 가시적 워터마크

가시적 워터마크는 수십 년간 사용되어 온 단순한 표시 방법으로, '초안'이나 '기밀'처럼 문서의 성격을 드러내는 데 주로 활용되어 왔습니다. 그러나 인공지능 분야에서 가시적 워터마크를 의무화하는 것은 현실에 맞지 않습니다. 쉽게 제거할 수 있어 실효성이 낮고, 오히려 근거 없는 안전감을 심어줄 수 있기 때문입니다.



집중 조명

인공지능 생성 텍스트

이미지 및 시청각 콘텐츠의 경우, 소비자가 해당 콘텐츠의 진위 여부나 인공지능 생성 여부를 확인할 수 있도록 하는 다양한 도구가 이미 존재합니다. 반면, 인공지능이 생성한 텍스트에 라벨링을 의무화하는 것은 현실적으로 여러 어려움을 수반합니다. 텍스트 형태의 인공지능 생성 콘텐츠에 대한 소비자 인식을 높이기 위해서는, 이용자가 인공지능 시스템과 상호작용하고 있음을 인지할 수 있도록 하는 요건 마련에 집중할 것을 권고합니다.

집중 조명

콘텐츠 출처 및 진위 확인을 위한 연합 (C2PA) 표준

C2PA 표준은 누구나 자신의 제품과 프로세스에 디지털 출처 정보를 손쉽게 적용할 수 있도록 설계된 공개 표준입니다. 여러 기술을 결합한 이 표준은 현재 국제표준화기구(ISO) 국제 표준으로의 승인을 앞두고 있습니다.

어떻게 작동하는가?

1. **파일 생성**

   이용자가 새 파일을 만들 때, 파일의 생성 방식과 편집 이력 등 출처 정보를 생성하는 도구를 활용할 수 있습니다.

2. **콘텐츠 자격증명**

   생성된 출처 정보는 콘텐츠 자격증명의 형태로 인코딩됩니다.

3. **암호화 서명**

   콘텐츠 자격증명은 파일을 생성한 소프트웨어 또는 하드웨어의 개인키로 서명되어 진위성이 담보되며, 이후 인증에 사용할 수 있도록 공개키도 함께 제공됩니다.

4. **내장 및 워터마킹**

   콘텐츠 자격증명은 파일 내부에 직접 내장하거나, 비가시성 워터마크·디지털 서명과 연결해 별도 데이터베이스에 저장하는 등 다양한 방식으로 보관할 수 있습니다.

5. **검증**

   파일의 진위성을 확인하고자 하는 이용자는 콘텐츠 자격증명을 검토하는 전용 도구를 통해 해당 파일이 원본인지, 수정된 것인지를 확인할 수 있습니다.

6. **표시**

   진위성이 확인된 콘텐츠에는 배지나 아이콘 등 명확한 시각적 표시를 붙일 수 있습니다.

콘텐츠 자격 증명이란 무엇인가?

최신 C2PA 표준은 복수의 기술을 결합한 콘텐츠 자격증명을 기반으로 합니다.

- 콘텐츠 자격증명에는 파일이 생성된 일시와 생성에 활용된 기술 등 해당 파일의 메타데이터가 담겨 있습니다.
- 메타데이터가 담긴 콘텐츠 자격증명에 워터마크 식별자와 디지털 서명을 추가로 결합할 수 있습니다.
- 이렇게 완성된 자격증명 패키지는 디지털 서명을 거쳐 해당 파일에 고유하게 귀속됩니다.
- 콘텐츠 자격증명은 파일 내부에 직접 내장되는 한편, 콘텐츠 자격증명 데이터베이스 등 외부 저장소에도 별도로 보관됩니다.

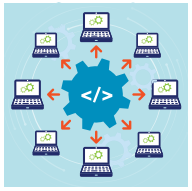
인공지능 공급망: 사업자 유형에 따른 역할과 책임

인공지능 공급망은 지속적으로 진화하고 있으며, 그 안에는 다양한 유형의 사업자가 참여하고 있습니다. 입법자들은 콘텐츠 진위성에 관한 정책을 수립할 때 이러한 사업자 유형의 차이를 반드시 고려해야 합니다. 사업자 유형에 따라 접근할 수 있는 정보의 범위가 다르고, 소비자 보호를 위해 취할 수 있는 조치도 달라지기 때문입니다.

예를 들어, 한 사업자가 인공지능 모델을 개발하면, 다른 사업자가 이를 애플리케이션에 통합하고, 또 다른 사업자가 그 애플리케이션을 활용해 새로운 콘텐츠를 생성할 수 있습니다.

모델 개발사업자

모델 개발사업자는 다양한 애플리케이션에 활용될 수 있는 인공지능 모델을 개발합니다. 예컨대 여러 용도에 맞게 응용할 수 있는 기반 모델을 구축하는 것이 대표적입니다. 하나의 기반 모델이 검색 엔진, 챗봇, 스팸 탐지 소프트웨어, 장문 텍스트 요약 도구 등 다양한 성격의 애플리케이션에 폭넓게 활용될 수 있습니다. 해당 사업자는 기반 모델의 개발 과정에 대한 정보를 보유하고 있으나, 다른 사업자가 해당 모델을 어떻게 배포하고 활용하는지에 대해서는 일반적으로 파악하기 어렵습니다.



통합사업자

통합사업자는 인공지능 모델을 특정 애플리케이션에 결합해 다른 사업자가 활용할 수 있는 형태로 제공합니다. 모델을 특정 서비스나 애플리케이션에 단순히 연결하는 데 그치는 경우도 있는 반면, 서비스나 애플리케이션에 탑재하기 전 모델을 미세 조정하거나 수정하는 경우도 있습니다. 이와 같은 사업자는 모델에 가한 변경 사항에 대한 정보는 파악하고 있으나, 모델의 최초 개발 과정이나 다른 사업자가 해당 인공지능 애플리케이션을 활용하는 구체적인 상황에 대해서는 직접적인 정보를 갖고 있지 않은 것이 일반적입니다. 예를 들어, 하나 이상의 인공지능 모델을 결합한 인공지능 애플리케이션을 개발하는 사업자가 통합사업자에 해당합니다.

이용사업자

인공지능 도구를 특정 목적에 맞게 활용하는 사업자를 흔히 이용사업자라고 합니다. 이용사업자는 특정 인공지능 기술을 언제, 어떻게 쓸지를 직접 결정하는 만큼 개별 활용 사례의 세부 사항을 잘 파악하고 있습니다. 다만, 이용사업자는 대개 다른 사업자로부터 인공지능 도구를 공급받기에, 해당 도구의 초기 학습 과정에 대한 직접적인 정보를 갖고 있지 않은 경우가 많습니다.

콘텐츠 진위성에 관한 모든 정책은 이러한 역할의 차이를 반드시 반영해야 합니다.

이처럼 성격이 다른 사업자들에게 획일적인 요건을 적용하는 정책은 소비자들에게 실질적인 도움이 되지 않습니다. 예를 들어, 콘텐츠를 생성하는 인공지능 애플리케이션을 개발한 사업자는 해당 애플리케이션이 만들어낸 콘텐츠에 출처 정보를 적용하기에 가장 적합한 위치에 있으나, 기반 모델의 개발사업자는 다른 사업자가 개발·활용하는 인공지능 애플리케이션을 통해 생성되는 콘텐츠에 워터마크나 기타 디지털 표시를 직접 적용할 기술적 수단을 갖추고 있기 어렵습니다.

법체계의 지속적 적합성 확보. 콘텐츠 출처 확인을 위한 최선의 기술적 해법은 시간이 지남에 따라 변화하기 마련입니다. 정책 입안자들은 이러한 변화를 수용할 수 있도록 입법 체계를 유연하게 설계해야 합니다. C2PA와 같은 개방형 표준을 적극 수용하는 것이 이러한 노력에 도움이 될 수 있습니다.

딥페이크와의 싸움

비가시성 워터마크, 디지털 지문, 보안 메타데이터와 같은 도구들은 콘텐츠의 진위성을 확인하는 데 도움이 될 수 있으며, 이를 통해 이용자는 자신이 접하는 이미지·영상·오디오 클립의 진위 여부를 판단할 수 있습니다.

이러한 도구들은 종합적으로 딥페이크 문제에 대응하는 데 기여합니다. 악의적 행위자가 자신의 딥페이크 콘텐츠에 스스로 허위임을 표시를 할 가능성은 없으므로, 이용자가 콘텐츠의 진위성을 스스로 확인할 수 있는 수단을 갖추는 것이 매우 중요합니다. 이를 위해서는 해당 도구들이 다양한 기기와 플랫폼에 걸쳐 폭넓게 도입되어야 합니다. 아울러 이용자들이 이러한 도구들을 제대로 이해하고 활용할 수 있도록 교육이 뒷받침되어야 합니다. 이를 통해 이용자들은 이미지나 영상을 접할 때 출처 정보를 확인하고, 온라인 정보 전반에 대해 건전한 비판적 시각을 갖출 수 있습니다.

악의적 행위자들은 인공지능을 포함한 기술을 기만적 목적으로 악용하는 새로운 방법을 계속해서 찾아낼 것입니다. 그러나 보안 메타데이터나 C2PA 표준과 같은 도구들은 선의의 행위자들이 자신의 콘텐츠 진위성을 입증하는 데 결정적인 역할을 할 수 있으며, 나아가 우리 모두가 온라인에서 사실과 허위를 가려내는 데 도움이 될 것입니다.

정책 입안자들은 딥페이크에 효과적으로 대응하는 방안을 검토할 때, 사업자 유형에 따라 수행하는 역할과 기능의 차이도 함께 고려해야 합니다. 서비스 수준, 기술적 특성, 기능, 이용자 기반 등 여러 측면에서 사업자마다 본질적인 차이가 있고, 그에 따라 대응 가능한 문제의 범위도 달라지기 때문입니다.

아울러 정책 입안자들은 서비스 유형별로 상이한 위험 수준을 인식할 필요가 있습니다. 예를 들어, 기업 간 (B2B) 소프트웨어 서비스는 소비자에게 직접 서비스를 제공하지 않고 이용자 기반도 제한적이라는 특성상, 이용자 안전이나 공공질서에 미치는 위험이 크지 않으며 딥페이크 관련 우려도 상대적으로 덜할 수 있습니다.

인공지능 정책, 소비자를 어떻게 도울 수 있나

정책 입안자들이 소비자가 온라인에서 사실과 허위를 가려낼 수 있도록 하려면, 콘텐츠의 인공지능 생성 여부를 식별할 수 있는 기존 도구들을 활용하는 것이 중요합니다.

정책목표	해결 방안
 **인공지능 시스템과 상호작용할 시 이용자에게 이를 알림**	사업자가 소비자와 직접 상호작용하는 인공지능 시스템을 제공할 때는, 명백히 인지 가능한 상황이 아닌 한 소비자에게 이를 고지해야 합니다.
 **소비자가 콘텐츠의 인공지능 생성 여부를 알 수 있도록 지원**	콘텐츠 출처 확인 및 인증 도구의 활용 의무는 소비자를 대상으로 하는 음성·이미지·영상 콘텐츠에 한정해야 하며, 텍스트나 기업 간 (B2B) 거래 환경에는 적용하지 않는 것이 바람직합니다. 소비자가 이메일 문구를 다듬거나 문서를 번역하는 데 인공지능을 쓰는 경우처럼, 텍스트에 콘텐츠 진위성 확인 체계를 적용하는 것은 현실적으로 맞지 않습니다. 기업 간 거래 환경에서는 각 사업자가 활용 목적에 맞는 콘텐츠 진위성 확인 방식을 자체적으로 마련할 수 있습니다.
 **인공지능 생성 콘텐츠 식별을 위한 글로벌 선도 기술의 활용을 지원**	인공지능 생성 콘텐츠 식별 의무는 C2PA와 같은 개방형 표준을 포함한 글로벌 선도 도구를 통해 이행할 수 있도록 해야 합니다. 현재 여러 나라가 콘텐츠 출처 확인 관련 규제를 검토하고 있는 와중, 국가마다 별도의 표준을 도입할 경우 소비자가 지속적으로 업데이트되는 글로벌 표준을 활용하는 데 걸림돌이 될 수 있습니다.
 **인공지능 생성 콘텐츠에 대한 투명성을 확보**	인공지능 생성 콘텐츠에 출처 정보를 표시할 의무는 원칙적으로 인공지능 애플리케이션 개발사업자에게 부과하는 것이 적절합니다. 콘텐츠를 실제로 만들어내는 주체가 인공지능 애플리케이션이기 때문입니다. 사업자는 기계 판독형 워터마크, 디지털 지문, 보안 메타데이터 등을 활용해 콘텐츠의 출처를 표시할 수 있습니다.
 **콘텐츠 출처 정보가 이용자에게 지속적으로 제공될 수 있도록 함**	보안상 불가피한 경우를 제외하고, 기계 판독형 워터마크·보안 메타데이터 등 콘텐츠 출처 정보를 무단으로 삭제하는 행위를 금지해야 합니다.