2026-10호

AI 산출물 신뢰성 확보를 위한 미국의 정책 접근 : 연방거래위원회 「AI 정확성 정책성명서(안)」 분석

주요국 AI·디지털 정책 모니터링 리포트

The LENS 2026

NIA 한국지능정보사회진흥원
NATIONAL INFORMATION SOCIETY AGENCY

목 차

1. AI 정확성 정책성명서(안) 분석 배경 및 개요 ····· 1
2. AI 정확성 정책성명서(안) 배경 ····· 2
3. AI 정확성 정책성명서(안) 주요 내용 ····· 4
4. AI 정확성 정책성명서(안)에 대한 공개 의견 ····· 6
5. AI 산출물 신뢰성 확보를 위한 주요국 정책 접근 ····· 9
6. AI 산출물 신뢰성 정책에 대한 함의 ····· 11

함께 보면 좋은 정책 자료 ····· 13
참고문헌 ····· 16

AI 산출물 신뢰성 확보를 위한 미국의 정책 접근
- 연방거래위원회 「AI 정확성 정책성명서」(안) 분석 -

1. AI 정확성 정책성명서(안) 분석 배경 및 개요

● **(AI 신뢰성 이행 논의 확대)** 글로벌 AI 윤리 논의에서 '신뢰할 수 있는 AI'는 공정성, 정확성, 투명성 등의 핵심 가치를 포괄하는 개념으로 제시되어 왔으며, 최근 이를 AI 개발-배포-이용 과정에서 구현하기 위한 이행 방안으로 논의가 확대되는 추세[1)]

- 초기에는 기업의 자율적 개발·운영이 강조되었으나, 생성형 AI 확산으로 산출물이 즉각 생성·활용되면서 AI 신뢰성 확보를 위한 구체적 이행·검증 수단의 중요성 부상
- 이에 각국은 AI 신뢰성 원칙을 법률, 지침, 자율 규범, 시험 도구 등으로 구체화하여 AI 시스템의 평가·관리와 기업의 책임 이행을 유도

● **(美 AI 정확성 정책성명서(안) 발표)** 연방거래위원회*가 '26년 7월 「AI 정확성 정책성명서(Policy Statement Concerning the Suppression of Accuracy in Artificial Intelligence Systems)」(안)**을 발표하면서, AI 산출물의 신뢰성 확보에 관한 미국의 정책 방향 제시[2)]

* Federal Trade Commission(FTC) : 반독점법 집행 및 소비자 보호를 주요 임무로 하는 연방 규제기관으로, 우리나라의 공정거래위원회와 유사 기능 수행

** '26년 7월 1일 공개된 의견 수렴용 초안으로, 제출된 공개 의견을 검토하여 최종 정책 마련 예정

- 본 성명서(안)은 미국 내에서 연방 차원의 포괄적 AI법이 부재한 가운데, 기존의 소비자 보호에 관한 법에 근거해 AI 산출물의 신뢰성 문제에 대응한 조치
- 주요국이 서로 다른 방식으로 AI 신뢰성 확보 수단을 마련하고 있어, 본 성명서(안)은 AI 산출물 신뢰성을 둘러싼 국가별 접근의 특징과 주요 쟁점을 파악하는 준거로 활용 가능

● **(분석 개요)** 본 보고서는 정책성명서(안)의 배경 및 주요 내용, 공개 의견, 주요국 정책과의 비교 분석을 종합하여 AI 산출물의 신뢰성 확보에 관한 정책적 함의 도출

- 먼저 트럼프 2기 행정부의 AI 정책 기조 속에서 정책성명서(안)이 마련되기까지 경과를 살펴보고, 이러한 정책 흐름의 연장선상에서 성명서(안)의 핵심 논리 구조와 강조점 분석
- 이어 성명서(안)의 전제 및 적용 방향에 대한 공개 의견*의 주요 입장을 파악하고, 주요국** 정책과의 비교를 통해 미국 접근의 특징을 도출한 후, 정책적 함의 제시

* 연방 규제정보 포털(regulations.gov)에 등록된 305개 의견 중 주제 무관, 중복, 비공개 의견을 제외한 188건 분석[3)] / ** 유럽·북미·아시아 각 권역의 AI 정책 선도국인 EU, 영국, 캐나다, 싱가포르 선정

2. AI 정확성 정책성명서(안) 배경

● **(AI 정책 전환과 정확성 의제 부상)** 트럼프 2기 행정부는 출범 직후 바이든 행정부의 AI 정책 기조를 폐기하고, AI 규제 장벽 제거와 이념적 편향 배제를 중심으로 정책 방향 전환

- '25년 1월 행정명령 제14148호 「유해한 행정명령 및 조치의 초기 철회」를 통해 바이든 행정부의 핵심 AI 정책인 행정명령 제14110호 「안전하고 신뢰할 수 있는 AI 개발 및 사용」을 폐지하며 기존 연방 AI 정책의 전환 공식화
- 이어 같은 달 발표한 행정명령 제14179호 「미국 AI 리더십의 장벽 제거」에서 글로벌 AI 우위 확보라는 목표하에 혁신을 저해하는 기존 정책·지침의 재검토를 지시하고, 이념적 편향이나 특정 사회적 의제가 개입되지 않은 AI 개발을 정책 원칙으로 제시
- 이러한 정책 기조는 △연방기관이 조달하는 AI에 진실 추구 및 이념적 중립성 원칙을 적용하는 **'연방 내부 관리'**, △AI 산출물에 영향을 미치는 주 규제에 대응하는 **'연방-주 관계 정비'**라는 두 축으로 구체화

[AI 정확성 정책성명서(안) 관련 정책 추진 경과]

● **(AI 객관성·중립성 원칙의 조달 관리)** ’25년 7월 발표된 「미국 AI 실행계획」에서는 AI가 객관적 진실을 추구하고 특정 이념에 편향되지 않아야 한다는 원칙을 명확히 하고, 행정명령*을 통해 이를 연방 AI 조달의 주요 기준으로 확립

* 제14319호 「연방정부의 이념 편향 AI 방지」 : ‘진실 추구(truth-seeking)’와 ‘이념적 중립성(ideological neutrality)’이라는 두 가지 ‘편향 없는 AI 원칙(unbiased AI principles)’을 명문화

- ’25년 12월 발표된 예산관리처 메모랜덤 M-26-04 「편향 없는 AI 원칙을 통한 AI에 대한 공공 신뢰 제고」는 신규 대규모 언어 모델 조달 계약에 편향 없는 AI 원칙 준수를 계약 요건으로 반영하고, 가능한 범위에서 기존 계약에도 동일한 요건을 반영하도록 요구
- 또한 각 기관에 조달 정책·절차를 개정하고, 기관 내 대규모 언어 모델 이용자가 편향 없는 AI 원칙에 위배되는 산출물을 보고할 수 있는 절차를 마련하도록 요구

● **(AI 산출물 정확성 관련 주(州) 규제 대응)** 주 AI 규제의 확산으로 규제 파편화와 연방 AI 정책 기조와의 충돌 가능성이 제기되자, 연방 차원의 정책 체계를 마련하고 AI 모델의 진실한 산출물 변경을 요구하는 주 규제에 대한 대응 기준 마련

- ’25년 12월 발표된 행정명령 제14365호 「AI에 대한 국가정책기초 확보」에서는 주 AI 규제와 관련해 △주별로 상이한 규제 환경에 따른 기업 부담, △AI 모델의 진실한 산출물 변경 요구, △연방 AI 정책과의 충돌 가능성을 지적
- 특히 차별적 영향 방지 규제가 형평성 등 다른 목적을 우선하게 함으로써 AI 모델의 진실한 산출물을 변경할 수 있다는 문제의식 제시

※ 정책성명서(안)에서는 특정 집단에 대한 차별적 취급 및 영향을 초래하는 AI 시스템 이용을 금지한 콜로라도주의 「AI법」(SB 24-205, 시행 전 ’26년 5월 「자동화의사결정기술법」으로 폐지·대체)을 대표 사례로 언급

- 이에 연방 AI 정책과 충돌하는 주법에 대한 효력 다툼 소송을 추진하는 한편, 주정부의 규제 내용과 집행 방침을 연방 보조금 지원 조건과 연계하는 방안 구체화

● **(AI 정확성 정책성명서(안) 발표)** 행정명령 제14365호 제7조에서 주(州)법에 따른 AI 산출물 조정과 「연방거래위원회법」상 기만행위 금지 규정의 관계를 설명하는 정책성명서 마련을 지시함에 따라, 연방거래위원회는 ’26년 7월 AI 정확성 정책성명서(안) 발표

- 정책성명서(안)은 새로운 규제를 마련하기보다, 기존 「연방거래위원회법」의 제5조의 소비자 기만 기준을 AI 산출물에 적용하는 방식 및 집행 방향을 제시
- 구체적으로 주법 준수나 이념적 목적 등 고지되지 않은 목적에 따라 AI 산출물을 조정할 경우, 소비자 기만 기준을 적용한다는 방침 표명
- 이에 따라 AI 산출물의 정확성은 기업의 기술적 자율적 관리 영역을 넘어, 연방법에 따른 조사 및 집행 영역으로 확대

3. AI 정확성 정책성명서(안) 주요 내용

3-1 핵심 논리 구조

● **(① 정확성에 대한 소비자 기대)** 연방거래위원회는 AI 기업이 소비자의 목적에 부합하는 최선의 산출물을 제공한다고 제시해 온 만큼, 소비자가 진실하고 정확한 산출물을 기대하는 것은 합리적이라고 전제

- 이러한 기대의 근거로 △기업이 AI를 문제 해결 도구로 제시해 왔다는 점과 △AI가 이용자의 문제 해결을 지원하는 도구라는 제품·서비스 자체의 성격을 제시
- 다만 AI 시스템은 간결성, 관련성, 명료성 등 여러 목적을 동시에 추구할 수 있으며, 이용자가 다른 목적을 명시적으로 요청한 경우에는 정확성이 우선되지 않을 수 있음

● **(② 정확성을 억제하는 산출물 조정)** AI 기업이 소비자가 요청하거나 합리적으로 기대하는 방향과 다르게 고지되지 않은 다른 목적*을 우선하여 AI 산출물을 의도적으로 변경·유도하는 경우, 「연방거래위원회법」상 기만 문제가 발생할 수 있다고 판단

* △'역사적 부정의' 시정을 위한 이념적 개입, △주(州)법상 책임 발생을 피하기 위한 '형평성(equity)' 우선, △정치적 논란이 될 만한 산출물 배제를 요구하는 대중적 압력, △기업·직원의 정치적 의제 추구 등을 제시

- 성명서(안)이 문제 삼는 것은 설계 결정에 따른 의도적 산출물 조정으로, 기술·자원상 한계에서 비롯된 오류(환각)는 해당하지 않으며, 기만행위 요건에 해당하면 이윤 추구, 여론 형성, 주(州)법 준수 등의 조정 동기는 면책 사유가 되지 않음

● **(③ 산출물 조정의 고지 여부·수준 판단)** 소비자의 요청·기대와 달리 특정 목적을 우선하도록 AI 시스템이 설계된 경우, 이를 명시적·가시적으로 고지했는지 고려

- 정확성 이외의 목표를 AI 시스템에 반영하는 것 자체보다 이를 소비자에게 충분히 알렸는지가 핵심으로, 관련 정보는 소비자가 기존 기대를 바꿀 수 있도록 명확하고 눈에 띄게(clear and conspicuous) 제시되어야 함

● **(④ 「연방거래위원회법」상 기만행위 판단)** 「연방거래위원회법」 제5조의 기만적 행위 또는 관행 금지 규정에 따라 AI 기업의 산출물 조정이 소비자 기만에 해당하는지 판단

※ 법 제5조 '기만'의 의미와 판단기준을 명확히 하기 위해 기존 판례의 원칙을 종합한 「기만에 관한 정책성명서(FTC Policy Statement on Deception」를 기준으로 적용

- △표시·누락 또는 행위가 소비자를 오인시킬 가능성이 있고, △이를 해당 상황에서 합리적으로 행동하는 소비자의 관점에서 판단하며, △그 내용이 제품·서비스 이용이나 구매 결정에 영향을 미칠 정도로 중요한 경우 기만행위 성립

[AI 정확성 정책성명서(안)의 핵심 논리 구조]

① 소비자 기대	소비자는 AI 시스템이 진실하고 정확한 산출물을 제공할 것으로 기대
⇓	⇓
② 산출물 조정	기업이 고지되지 않은 목적을 우선해 AI 시스템의 산출물 조정
⇓	⇓
③ 고지 수준 판단	해당 조정 방식을 명확하고 눈에 띄게 고지했는지 판단
⇓	⇓
④ 기만행위 판단	「연방거래위원회법」 제5조상 기만행위 여부 판단

3-2 정책적 지향점

- **(① 정확성 중심의 가치 설정)** 진실 추구 및 이념적 중립성을 강조해 온 기존 정책 기조에 따라 정확성을 소비자가 AI에 기본적으로 기대하는 목적으로 전제함으로써, 여러 AI 신뢰성 가치 가운데 정확성에 상대적으로 우선적인 지위 부여

  - 일반적인 AI 신뢰성 논의에서는 정확성, 공정성, 형평성 등을 함께 충족해야 할 가치로 보지만, 성명서(안)은 정확성을 우선하고 다른 가치를 정확성을 낮출 수 있는 '다른 목적'으로 봄(6페이지 'AI 신뢰성 체계에서 '정확성'의 의미' 참조)

- **(② 편향 없는 AI 원칙의 민간 확장)** 연방 조달에 적용하던 편향 없는 AI 원칙을 소비자 보호에 관한 법률인 「연방거래위원회법」을 통해 민간 영역의 AI 시스템에도 간접적으로 적용하려는 방향 시사

  - 연방 조달에서는 이념적 중립성을 AI 시스템 설계에 반영해야 할 요건으로 제시하였으나, 민간 영역의 AI 시스템에는 고지를 정책 수단으로 활용함으로써, 모델 설계에 대한 규제 부담을 최소화하면서 동일한 정책 기조 반영

- **(③ 연방 중심의 규제 통일성 추구)** 주(州)법 준수를 제5조 적용의 면책 사유로 인정하지 않고 충돌하는 주법에 대한 묵시적 선점 가능성을 제시한다는 점에서, 주별 AI 규제를 견제하고 연방 기준의 통일성을 확보하려는 성격을 가짐

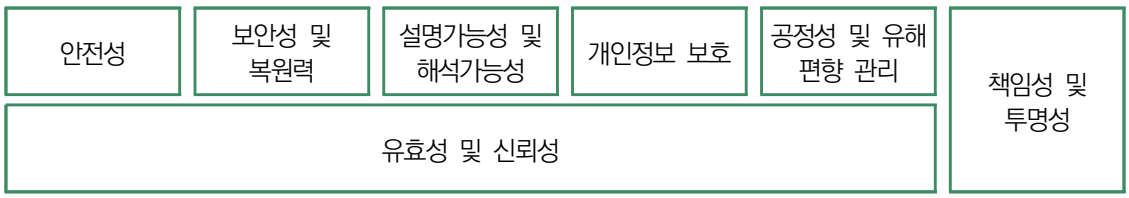
  - 정책성명서 자체가 주법을 무효화하지는 않지만, 주법에 따른 AI 산출물 조정도 연방 차원의 집행 대상이 될 수 있다는 정책 방향을 명확히 함

  - 통일된 연방 입법이 마련되기 전까지 연방거래위원회의 집행 권한을 활용하여, 소비자 보호와 동시에 연방 중심의 규제 일관성을 뒷받침하려는 접근

참고 AI 신뢰성 체계에서 '정확성'의 의미[11]

- 정책성명서(안)은 정확성을 공정성, 형평성 등 다른 목적에 우선하는 기준으로 삼고 있으나, 기존 AI 신뢰성 체계는 정확성을 신뢰성을 구성하는 여러 특성 중 하나로 규정
  - 미국에서 이를 구체적으로 제시한 문서는 국립표준기술연구소(NIST)의 「AI 위험관리 프레임워크(AI Risk Management Framework)」로, 정확성을 ISO/IEC 기술 규격서에 따라 "관측·계산·추정의 결과가 참값 또는 참으로 인정되는 값에 근접한 정도"로 정의
  - AI RMF는 신뢰할 수 있는 AI의 7가지 특징 중 정확성을 독립된 특성이 아니라 '유효성 및 신뢰성(valid and reliable)'을 구성하는 하위 지표로 규정하고, 신뢰할 수 있는 AI의 필수 조건이자 공정성, 안전성 등 다른 특성을 뒷받침하는 기반으로 제시

[「AI 위험관리 프레임워크」 中 신뢰할 수 있는 AI 시스템의 특징]



4. AI 정확성 정책성명서(안)에 대한 공개 의견

- **(의견 종합)** 의견 분석* 결과, 정책성명서(안)에 대한 입장은 일률적인 찬반으로 구분되기보다 동일한 제출 의견에서도 쟁점에 따라 동의·비동의·보완 요구가 함께 제시됨

  * '26년 8월 31일 기준 등록된 의견 305건 중 주제와 무관한 의견 111건, 중복 의견 5건, 비공개 의견 1건을 제외한 188건을 분석 대상으로 선정. 제출자는 개인 122건, 단체 66건으로 구성됨

  - 정책성명서(안)의 기본 전제에 동의하면서도 적용 범위에는 비동의하거나, 기본 전제에는 동의하지 않으면서도 소비자 오인 방지와 정보고지 필요성은 인정하는 등 하나의 제출 의견에서도 쟁점별로 상이한 입장이 제시됨
  - 특히 소비자 기대, 기만행위 법리 적용, 고지 의무, 편향·안전 조치의 성격, 편집적 판단 및 표현의 자유, 주법과 연방 권한의 관계 및 정책성명서(안)의 법적 성격 등을 중심으로 입장 차이와 보완 요구가 집중됨

- **(기만행위 판단 관련 쟁점)** 소비자 기대, 기만행위의 판단 범위, 편향 보정 및 안전 조치와 기만적 산출물 조정의 구별을 중심으로 성명서(안)의 전제와 적용 방식에 대한 의견 제기
  - **AI 정확성에 대한 소비자 기대** : 정확성을 표방하거나 사실 확인·정보 제공을 주된 용도로 하는 AI에서는 정확성에 대한 기대가 합리적이라는 의견과, 서비스의 용도와 기업의 성능 주장 등에 따라 기대 수준이 달라질 수 있다는 의견 제시

- **기만행위 법리 적용** : 기업이 표방한 목적과 실제 설계 목적 간 중요한 차이를 숨긴 경우 「연방거래위원회법」 제5조를 적용할 수 있다는 견해와, 모델의 설계 목적이나 출력까지 판단 대상으로 확대하는 것은 과도하다는 지적이 제기됨

- **편향 보정 행위의 성격** : 편향 보정은 데이터의 편중을 바로잡아 정확성을 높일 수도 있어 그 자체를 정확성 억제로 보기 어렵고, 보정의 목적과 기준 및 고지 여부를 함께 고려할 필요가 있다는 의견이 제시됨

- **안전 조치 행위의 성격** : 안전 조치 역시 산출물 조정을 수반하므로 기만적 조정과의 구별 기준이 필요하며, 정당한 안전 및 법 준수 목적의 조정에 적용할 예외 범위를 명확히 할 필요가 있다는 요구가 제기됨

● **(적용·권한 관련 쟁점)** 공시를 통한 기만 방지의 실효성과 정부 개입의 범위, 주법과 연방법의 관계 및 정책성명서(안)의 법적 성격 등을 둘러싸고 다양한 입장 제기

- **공시 의무** : 이용자가 실제 판단하는 시점에 핵심 정보를 알기 쉽게 고지해야 한다는 데 공감하면서도, 충분한 공시의 수준, 시점, 재고지 범위 등을 구체화할 필요성이 강조됨

- **편집적 판단 및 표현의 자유** : 모델의 학습 자료 선정, 가드레일, 출력 정책 등에 대한 정부 개입이 기업의 편집적·가치적 판단과 충돌할 수 있어, 상업적 표시와 실제 작동 간 차이에 판단 대상을 한정해야 한다는 견해 제시

- **주법과 연방 권한 간 관계** : 주법과 「연방거래위원회법」 제5조가 실제로 충돌하는 경우에 한하여 선점 여부를 판단해야 하며, 연방 차원의 대체 보호 없이 주법을 제한해서는 안 된다는 지적

- **정책성명서(안)의 성격 및 정당성** : 정책성명서(안)은 「연방거래위원회법」 제5조의 집행 방향을 설명하는 비구속적 지침으로, 주법을 대체하거나 새로운 준수 의무를 설정할 법적 효력이 없다는 지적과 함께 판단 요소와 집행 기준을 구체화할 필요가 있다는 요구 제기

[AI 정확성 정책성명서(안)에 대한 쟁점별 의견]

※ [동의]는 정책성명서(안)의 전제·적용 방향에 부합하는 의견, [비동의]는 반론·한계 지적, [보완]은 기준·예외·절차 등의 구체화 및 개선 등에 대한 요구를 의미. 각 표시는 의견 제출자의 전체 입장이 아닌 해당 쟁점에 관한 개별 의견의 주된 성격에 따라 구분

쟁점 사안	주요 의견
AI 정확성에 대한 소비자 기대	▪ [동의] 객관성 및 정확성을 표방하거나 사실 확인 및 정보 제공을 주된 용도로 하는 AI에서는 정확성에 대한 소비자 기대가 합리적이며 제품 선택에도 중요 ▪ [비동의] 생성형 AI 출력에는 판단과 해석이 포함되어 단일한 정답을 설정하기 불가능하며, 소비자가 정확성을 항상 최우선으로 기대한다는 실증적 근거 부족 ▪ [보완] 서비스의 용도, 기업의 구체적인 성능 주장, 홍보 문구의 명확성 등을 고려하여 소비자 기대를 판단할 필요

쟁점 사안	주요 의견
기만행위 법리 적용	▪ [동의] 목적에 관계 없이 기업이 표방한 목적과 실제 설계 목적 간 중요한 차이를 숨겼다면 기만행위 법리를 적용할 수 있으며, 개별 출력의 오류보다 설계 목적의 은폐가 핵심 ▪ [비동의] 허위·오인 표시와 중요 정보 누락은 현행 법리로 대응할 수 있으며, 모델의 관점·목적·출력까지 판단 대상에 포함하면 기만행위의 범위가 과도하게 넓어질 우려 ▪ [보완] 정확성 저하 자체보다, 기만행위의 세 가지 요건 중 소비자 오인 가능성과 의사결정에 미친 영향을 중심으로 판단할 필요
공시 의무	▪ [동의] 이용약관의 면책 문구만으로는 충분하지 않으며, 이용자가 실제로 판단하는 시점에 핵심 정보를 알기 쉬운 형태로 고지할 필요 ▪ [비동의] 공시가 실질적인 위험이나 차별적 설계를 정당화하는 면책 근거로 작용할 우려 ▪ [보완] 배포 이후 설정이나 출력 정책이 실질적으로 변경되면 재고지하되, 서비스의 이용 맥락과 위험 수준에 비례한 방식을 적용할 필요 ▪ [보완] 충분한 공시의 기준이 불분명하여 사업자와 이용자의 예측 가능성이 낮아지므로, 고지 시점과 학습·설계 방식의 고지 범위를 구체화할 필요
편향 보정 행위의 성격	▪ [비동의] 편향 보정은 학습 데이터의 사회적 편중을 바로잡아 정확성을 높일 수도 있으므로, 그 자체를 정확성을 저하한 행위로 단정하기 어려움 ▪ [비동의] 편향 보정을 기만행위로 간주하면 보정한 기업만 법적 위험을 부담하고, 편향을 시정하지 않는 기업이 상대적으로 유리해질 수 있음 ▪ [보완] 편향 보정의 기만행위 해당 여부는 정치적 성향과 관계없이, 보정의 목적과 기준을 충분히 고지했는지를 중심으로 판단할 필요
안전 조치 행위의 성격	▪ [보완] 안전 조치 역시 출력을 조정하는 설계 행위여서 기만적 유도와의 객관적 구별 기준을 세우기 어렵고, 개별 예외 열거 방식은 실제 개발·운영 과정을 반영하기 어려움 ▪ [보완] 아동 보호 및 사기 방지 등 정당한 안전 조치는 기만행위와 구별하고, 안전 및 법 준수 목적의 출력 제한에 적용할 예외 기준을 마련할 필요 ▪ [보완] 안전 조치의 목적과 적용 범위를 문서화하고 이용자에게 고지할 필요
편집적 판단 및 표현의 자유	▪ [비동의] 학습 자료 선정, 가드레일, 출력 정책은 헌법상 보호되는 기업의 편집적 판단에 해당하여 정부의 규율 대상으로 삼기 어려움 ▪ [비동의] 정부가 출력의 정확성을 직접 판단하면 내용·관점 규제로 이어질 수 있고, 객관적 사실이 아닌 내용의 공시를 강제하면 표현의 자유를 제약할 우려 ▪ [보완] 판단 대상을 출력 내용이 아니라 기업의 상업적 표시와 실제 작동 간 차이로 한정하고, 사실에 관한 성능 주장과 기업의 편집적·가치적 판단을 구별할 필요
주법과 연방 권한 간 관계	▪ [비동의] 선점 여부는 행정명령이 아니라 연방법과 기존 선점 법리에 따라 판단해야 하며, 소비자 보호와 차별 금지는 전통적으로 주정부가 규율해 온 영역 ▪ [비동의] 연방 차원의 대체 보호 없이 주법만 제한하면 소비자 보호의 공백이 발생할 우려 ▪ [보완] 주법에 따른 출력 조정 사실을 고지하면 제5조와 동시 준수할 수 있으므로, 양자가 실제로 충돌하는 경우에 한해 선점 여부를 판단할 필요
정책성명서의 성격 및 정당성	▪ [비동의] 전국적인 통일 기준은 행정기관의 지침이 아니라 의회 입법으로 마련하는 것이 적절 ▪ [비동의] 정책성명서에는 주법을 대체하거나 새로운 준수 기준을 설정할 법적 효력이 없어, 실질적인 의무를 부과하려면 정식 규칙 제정 절차가 필요 ▪ [보완] 정책성명서가 제5조의 집행 방향을 설명하는 비구속적 지침임을 명확히 하고, 기만행위 판단과 주법 선점 문제를 구분하여 다룰 필요 ▪ [보완] 구체적인 판단 요소와 집행 사례를 제시하여 예측 가능성을 높이고 자의적인 적용을 방지할 필요

5. AI 산출물 신뢰성 확보를 위한 주요국 정책 접근

● **(비교 개관)** 정확성 확보 수단, 안전·편향 관련 조치, 정보고지를 중심으로 주요국 정책을 비교한 결과, 미국과 비교 대상국 모두 AI 산출물의 신뢰성 확보를 지향

- 다만 각국은 공공·민간 등 AI 이용 영역과 위험 수준에 따라 신뢰성 확보 기준을 달리하고, 법적 의무와 자율 규범을 선택적으로 또는 병행하여 활용

● **(① 정확성 확보 수단)** 미국과 주요국 모두 모든 AI에 일률적인 정확성 수준을 요구하지 않고, 적용 대상과 이용 목적, 오류가 미치는 영향에 따라 요구 수준을 달리 설정

- 미국은 연방 조달 대규모 언어 모델에는 진실성과 객관성을 시스템 요건으로 부과하는 반면, 민간 시장에서 제공되는 AI 시스템(이하 '민간 AI')에는 정확성 외의 목적에 따른 산출물 조정과 그 공시 여부에 초점

- EU는 고위험 AI에 정확성·강건성의 성능 요건을 법정 의무로 부과하는 반면, 영국·캐나다·싱가포르는 이용 맥락에 따른 시험 및 평가와 지속적인 품질관리에 중점

● **(② 안전·편향 관련 조치)** 미국과 비교 대상국은 안전 확보 및 편향 완화 조치에 부여하는 정책적 중요성과, 이들 조치와 정확성 확보 간의 관계 설정에서 차이를 보임

- 미국의 연방 조달에서는 안전 필터 및 편향 평가 등을 진실 추구와 이념적 중립성 원칙과의 부합 여부를 확인하는 보완적인 요소로 활용하고, 민간 AI에서는 정확성 외의 목적을 위한 산출물 조정을 고지하지 않은 경우를 문제 삼음

- 비교 대상국에서는 안전성 확보와 편향 완화를 정확성과 상충되는 목적이 아니라, AI 신뢰성을 담보하기 위해 정확성과 병행하여 관리·충족해야 할 핵심적 동반 요소로 설정

- 다만 영국은 채용·신용평가 등 개인정보 처리 AI에 대해 편향 완화가 전체 정확도를 낮출 수 있음을 인정하되, 그 조정 방식을 개발 초기에 정하여 문서로 남기도록 요구

● **(③ 정보고지)** 미국과 비교 대상국 모두 이용자나 정책당국이 AI 시스템과 산출물을 판단할 수 있도록 관련 정보의 제출 또는 고지를 요구하거나 권고

- 미국은 연방 조달에서는 관련 자료를 정부에 제출하도록 요구하고, 민간 AI에서는 고지 내용의 상세함보다 소비자가 인지할 수 있도록 명확하고 눈에 띄게 고지했는지를 중시

- EU는 고위험 AI의 이용 목적, 정확성 지표, 위험, 한계를 배포자에게 제공하도록 의무화하고, 영국은 위험 비례적 정보 제공 원칙과 공공부문 알고리즘 고지 의무를 병행

- 캐나다는 연방의 자동결정에 관한 고지·설명과 알고리즘 영향 평가 결과 고지를 의무화하는 반면 민간 생성형 AI에는 자율 고지를 권고하며, 싱가포르는 생성형 AI의 평가 결과, 안전 조치, 한계의 고지를 자율원칙으로 제시

[주요국 AI 산출물 신뢰성 정책 비교]

국가	정책 수단	① 정확성 확보 수단	② 안전·편향 관련 조치	③ 정보고지
미국	행정명령[12], 조달지침[13]	▪ [연방 조달 LLM] 계약조건으로서, 사실·분석 질의에 진실성과 객관성을 추구하고 정보가 불완전한 경우 불확실성 인정을 요구	▪ [연방 조달 LLM] 안전 필터 및 편향 평가 등은 진실 추구와 이념적 중립성 원칙과의 부합 여부를 확인하는 요소로 활용	▪ [연방 조달 LLM] 공급기업에 모델·시스템·데이터 카드 등 기본 자료 요청, 필요시 평가 결과 및 안전 필터 등 추가 자료 요청
	기존 법령[14], 정책성명서(안)[15]	▪ [민간 AI] 기술적 한계에 따른 오류와 구분하여 비고지 목적에 따른 의도적 정확성 억제만을 주된 기만행위 판단 대상으로 설정	▪ [민간 AI] 안전·편향 조정 자체는 금지하지 않으나, 이를 고지하지 않은 채 정확성보다 우선하면 기만에 해당할 수 있다고 판단	▪ [민간 AI] 정확성 외의 다른 목적을 우선하는 경우 이를 명확하고 눈에 띄게 고지했는지를 기만행위 판단 요소로 고려
EU	AI 특화 법령[16]	▪ [고위험 AI] 법정 의무로서, 의도된 목적에 부합하는 정확성·강건성 수준을 확보하고 관련 지표를 제시하도록 요구	▪ [고위험 AI] 데이터 편향의 검토·완화를 정확성과 동시에 충족해야 할 법정 의무로 규정 - 편향의 탐지 및 교정을 위해 특별한 범주의 개인정보*에 대한 예외적 처리 허용 * 인종·건강·정치적 견해 등	▪ [고위험 AI] AI 제공자가 배포자에게 제공하는 사용설명서에 이용 목적, 정확성 지표, 성능 한계 및 예견가능한 위험 기재 요구
영국	비구속적 정책 원칙[17]	▪ [AI 전반] 공통의 정량적 기준을 두지 않고, 이용 목적과 오류가 미치는 영향에 따라 규제기관이 정확성 수준을 판단할 것을 요구	▪ [AI 전반] 안전, 보안, 강건성, 공정성을 공통 원칙으로 제시하고, 규제기관이 이용 맥락에 따른 위험 수준에 비례하여 적용	▪ [AI 전반] AI 이용 사실, 목적, 작동 방식, 한계에 관한 정보를 이용 맥락에 따른 위험 수준에 비례하여 제공
	규제기관 지침[18][19]	▪ [개인정보 처리 AI] 사람에 관한 추론을 하는 AI에 대해 이용 목적에 충분한 수준의 정확성 요구	▪ [개인정보 처리 AI] 편향을 줄이는 조치가 전체 정확도를 낮출 수 있음을 인정하되, 조정 방식을 개발 초기에 정하여 문서로 남기도록 요구	
	공공부문 의무 표준[20]			▪ [공공부문 AI] 공적 의사결정에 중대한 영향을 미치거나 국민과 직접 상호작용하는 알고리즘 도구에 대하여 고지 의무화
캐나다	연방 행정지침[21]	▪ [연방 자동결정] 행정 내부 규범으로서, 알고리즘 영향 평가로 산정한 영향 수준에 따라 배포 전 시험을 거치고 배포 후 정확성과 부작용을 지속 모니터링하도록 요구	▪ [연방 자동결정] 의도하지 않은 데이터 편향에 대한 시험을 요구하고, 배포 후 공정성을 정확성과 함께 지속 모니터링하도록 요구	▪ [연방 자동결정] 자동결정 이용 사실을 사전에 고지하고 결정의 방식·이유를 설명하며, 알고리즘 영향평가 결과를 고지하도록 요구
	민간 자율 규범[22]	▪ [민간 생성형 AI] 법적 기준을 두지 않고, 배포 전 다양한 과업·맥락에서의 시험과 취약점 발견을 위한 적대적 시험을 자율원칙으로 제시	▪ [민간 생성형 AI] 학습 데이터의 품질과 편향을 관리하고, 배포 전 편향된 산출물의 위험을 평가 및 완화하도록 자율원칙으로 제시	▪ [민간 생성형 AI] 시스템의 능력과 한계, 학습 데이터의 유형 및 위험 식별·완화 조치를 고지하도록 권고
싱가포르	자율 프레임워크[23]	▪ [생성형 AI] 정확성의 최저 기준을 두지 않고, 사실성을 표준 안전 평가 항목의 하나로 포함 - 고위험 용도 AI의 정확성 기준은 업계와 분야별 정책 당국이 함께 정하도록 권고	▪ [생성형 AI] 벤치마킹과 레드팀 시험을 주된 평가 방법으로 제시하고, 편향성과 유해성을 사실성 및 강건성과 함께 공통 평가 항목에 포함 - 이용 맥락별 위험평가를 거쳐 미세조정, 입출력 필터 등을 통한 유해 산출 완화 권고	▪ [생성형 AI] 학습 데이터의 유형·처리 방식, 평가 결과, 안전·완화 조치, 알려진 위험·한계 및 의도된 이용 목적을 고지하도록 권고

6. AI 산출물 신뢰성 정책에 대한 함의

이번 정책성명서(안)이 AI 산출물의 신뢰성을 실제 정책과 제도로 이행해 나가는 과정에서 어떠한 쟁점이 제기되며, 향후 관련 정책에 어떠한 함의를 갖는지를 중심으로 살펴봄

- **(신뢰성 요소 간 우선순위)** 공정성, 안전성, 정확성, 투명성 등 AI 신뢰성의 여러 요소가 충돌할 수 있는 상황에서, 무엇을 우선할지 그리고 그 우선순위를 누가 결정할지에 대한 논의가 확대될 가능성
  - AI 신뢰성 원칙과 기준을 실제 정책과 시스템 설계에 적용하는 과정에서는 어떤 가치를 어느 수준까지 우선할 것인지에 대한 판단이 불가피하며, 문화적·정치적 맥락에 따라 그 기준과 정당성을 둘러싼 이견이 뒤따르기 쉬움
  - 이번 정책성명서(안) 역시 정확성보다 다른 목적을 우선한 산출물 조정을 소비자 기만의 문제로 다룸으로써, 신뢰성 요소 간 우선순위가 법·정책의 선택으로 구체화된 사례
  - 상이한 국가별 규범과 기업의 설계 선택을 어떻게 조정할 것인지, 글로벌 기준과 국가·기업 차원의 판단 사이에서 어떤 가치를 우선할 것인지의 문제가 제기될 수 있음
- **(공지 중심 접근의 조건)** AI 산출물의 신뢰성을 공지와 이용자 선택에 상당 부분 맡기는 접근이 실효성을 가지려면, 이용자가 공지된 정보를 이해·검증하고 대체 서비스로 옮겨갈 수 있는 조건이 어디까지 갖춰져 있는지가 관건
  - 미국의 접근은 정확성 외의 설계 목적을 충분히 고지하면 이용자가 이를 판단해 서비스를 선택할 수 있다는 점에서, 신뢰성 확보의 상당 부분을 정보고지와 시장 선택에 의존
  - 위험이 낮고 서비스 간 성능 비교가 용이한 영역에서는 공시가 이용자의 선택을 뒷받침할 수 있으나, 의료·고용 등 오류의 영향이 크거나 개별 산출물의 정확성을 확인하기 어려운 영역에서는 공시만으로 합리적 선택을 기대하기 어려움
  - 이 조건이 갖춰지지 않은 상태에서는 공시가 이용자의 판단을 돕기보다 기업이 책임을 면하는 근거로 작용할 수 있어, 영역에 따라 공시 외의 수단의 병행을 검토할 필요
- **(신뢰성 판단 주체와 책임)** AI 산출물의 정확성 및 편향 여부를 누가 판단하고, 그 판단에 따른 책임을 어떻게 배분할 것인지가 신뢰성 정책에서 중요한 쟁점이 될 수 있음
  - 이번 정책성명서(안)은 산출물 조정에 관한 판단과 공시 책무는 기업에, 고지된 정보에 따른 서비스 선택은 이용자에게, 기만 여부에 대한 최종 판단은 정부에 두는 구조

- 어느 주체에 더 큰 책임을 둘지에 따라 자율성, 선택권, 정부 감독 측면에서 기대되는 효과와 한계가 달라질 수 있어, 역할과 책임의 적정한 균형을 찾는 것이 중요

- 비교 대상국 역시 정부, 기업, 이용자에게 판단과 책임을 배분하는 방식이 서로 다른 만큼, 향후 정책 시행 과정에서 나타나는 효과와 한계를 바탕으로 정책 수단을 지속적으로 조정·보완할 필요

● **(글로벌 신뢰성 협력 및 거버넌스)** 국가마다 신뢰성을 구성하는 요소와 이행 수단이 다른 상황에서, 국제적 정합성을 확보할 영역과 각국의 정책적 자율성으로 남겨둘 범위를 조율하는 것이 주요 과제로 대두

- 미국 및 비교 대상국은 신뢰성 요소 간의 우선순위와 감독, 공시, 시험 등 이행 수단 간의 차이를 보이며 동일한 개념을 제도적으로 다르게 구현

- 한편, OECD 등 다자 기구를 통한 원칙 형성에서 나아가, AI 안전연구소 간 공동 평가, 국제표준 마련, 위험·사고 정보 공유 등 검증 기반을 넓히는 협력도 확대되는 추세

- 이에 향후 국제 협력은 가치와 원칙의 통일에 그치지 않고, 시험 방법, 평가지표, 사고 기록 등 상호 비교·검증을 돕는 공통 기반을 모색하는 방향으로도 전개될 가능성

□ 함께 보면 좋은 정책 자료

자료명	생성형 AI 서비스 관리 잠정방법 生成式人工智能服务管理暂行办法[24]				
국가/기관	중국	자료 유형	법령·제도	발표 일자	2023. 7. 13.

- 중국은 '23년 7월 발표한 「생성형 AI 서비스 관리 잠정방법(生成式人工智能服务管理暂行办法)」에서 학습데이터, 데이터 라벨링, 생성 콘텐츠 등 단계별로 정확성 관련 의무를 규정
  - **(학습데이터: 제7조 4호)** 학습데이터의 품질을 개선하고 진실성·정확성·객관성·다양성을 향상시키기 위한 효과적인 조치를 취해야 함
  - **(데이터 라벨링: 제8조)** 데이터 라벨링 작업을 수행할 때, 제공자는 명확하고 실행 가능한 규칙을 수립하고, 데이터 라벨링 품질 평가 및 정확성 표본 검증을 실시해야 함
  - **(생성 콘텐츠: 제4조 5호)** 서비스 유형의 특성에 따라 생성형 AI 서비스의 투명성을 향상시키고, 생성 콘텐츠의 정확성과 신뢰성을 높이기 위한 효과적인 조치를 취해야 함
  - **(사후관리: 제14조)** 불법·부적절한 콘텐츠 발견 시 생성·전송 중단 및 삭제 등의 조치를 취하고, 모델 최적화·훈련 등을 통해 지속적으로 수정해야 함
  - **(안전평가·감독: 제17조·제19조)** 제공자는 여론 형성 또는 사회동원 기능을 갖춘 생성형 AI 서비스에 대한 안전평가 및 알고리즘 등록을 실시하며, 감독 과정에서 훈련데이터 출처·규모·유형, 라벨링 규칙, 알고리즘 메커니즘 등에 관한 설명과 필요한 기술·데이터 지원 및 협조

자료명	AI 기반 의인화된 상호작용 서비스 관리 잠정방법 人工智能拟人化互动服务管理暂行办法[25]				
국가/기관	중국	자료 유형	법령·제도	발표 일자	2026. 4. 10.

- 중국은 '26년 4월 「AI 기반 의인화된 상호작용 서비스 관리 잠정방법(人工智能拟人化互动服务管理暂行办法)」를 제정하며, 의인화된 AI 서비스 제공자가 감정조작 등을 통해 사용자의 불합리한 결정을 유도하여 정당한 권익을 침해하는 행위를 금지
  - **(학습데이터: 제11조)** 학습데이터의 투명성·신뢰성을 높이기 위해 데이터 정제·라벨링 등을 강화하고, 데이터를 정기적으로 점검·최적화하여 서비스 성능을 지속적으로 향상하도록 규정
  - **(편차·성능 관리: 제10조)** AI 서비스 전 생애주기에 걸쳐 안전 모니터링 및 위험평가를 실시하고, 시스템 편향을 적시에 발견·교정하도록 규정
  - **(안전평가: 제22·23조)** 신규 서비스 출시·중대 서비스 변경·등록 사용자 100만 명 이상 또는 월간 활성 사용자 10만 명 등 특정 조건을 충족하는 서비스 제공자는 안전평가를 실시해야 하며, 평가 시 훈련데이터 처리 및 중대한 안전위험의 시정 상황 등을 점검해야 함

자료명	AI 사업자 가이드라인 AI事業者ガイドライン[26]				
국가/기관	일본	자료 유형	지침·가이드라인	발표 일자	2024. 4. 19.

● 일본 경제산업성(経済産業省)·총무성(総務省)은 '24년 4월 'AI 사업자 가이드라인(AI事業者ガイドライン)'을 발표하며 학습데이터의 정확성·최신성과 AI 출력의 정확성·신뢰성을 강조하고, 개발자·제공자·사용자 등 각 역할에 따른 정확성 관리방안을 제시

- **(공통 원칙: 인간중심적)** 모든 AI 사업자는 일본 헌법이 보장하거나 국제적으로 인정된 인권을 침해하지 않는 방식으로 서비스를 개발·제공·이용해야 하며, 개인의 권리·이익에 중대한 영향을 미칠 수 있는 분야에서 AI 프로파일링을 수행할 경우 출력의 정확성을 최대한 유지해야 함
- **(공통 원칙: 안전)** 모든 AI 사업자는 AI 시스템·서비스가 요구사항에 충분히 부합하도록 작동하는지 확인하는 과정에서 출력의 정확성을 확보하는 등 신뢰성을 높이고, AI 시스템·서비스의 특성과 목적에 따라 학습데이터의 정확성·최신성(데이터 적합성)을 보장해야 함
- **(AI 개발자)** 정확성에 지나치게 집중할 경우 개인정보보호·공정성 등 다른 가치가 훼손될 수 있음을 고려해 경영상 위험과 사회적 영향을 종합적으로 판단하고 필요한 수정조치를 취해야 함
- **(AI 제공자)** AI 시스템 구현 시 제공 시점을 기준으로 AI 시스템·서비스의 정확성과 필요한 경우 훈련데이터의 최신성(데이터 적합성)을 확보해야 함
- **(AI 사용자)** AI 시스템 사용 시 AI 출력의 정확도 및 위험 수준을 이해하고, 다양한 위험 요인을 확인한 후 AI 출력을 활용해야 함

자료명	인공지능(AI) 제품 및 시스템 평가 참고지침 人工智慧(AI)產品與系統評測參考指引(草案)[27]				
국가/기관	대만	자료 유형	지침·가이드라인	발표 일자	2024. 3. 29.

● 대만은 '24년 3월 '인공지능(AI) 제품 및 시스템 평가 참고지침 (人工智慧產品與系統評測參考指引)' 초안을 발표하며 AI 시스템의 정확성을 독립적인 평가항목으로 설정하고, 모델의 저적합·과적합을 줄이기 위한 평가 및 관리방안을 제시

- 해당 지침은 고위험 AI 제품·시스템에 대해 정확성, 안전성, 설명가능성, 복원력, 공정성, 투명, 책임성, 신뢰성, 개인정보, 보안 등 10개 평가항목을 제시하고, 실제 환경에서의 테스트를 통한 시스템 검증 방안을 제시

  ※ 고위험 AI 시스템 : 중요 기반시설, 장거리 생체인식 시스템, 공공서비스 및 민간 중요 서비스, 출입국·망명·국경관리 등

- **(정확성 평가)** AI 제품·시스템의 출력 결과와 실제 결과 간의 근접성을 측정하는 것으로 평가지표 선정과 모델 학습 과정에서 저적합 및 과적합을 줄이도록 권고
- **(지속적 운영관리)** AI 시스템의 위험을 정기적으로 평가하고 모니터링·검토 절차를 마련하는 등 AI 시스템의 위험 변화에 대응하도록 권고

자료명	자발적 AI 안전표준 Voluntary AI Safety Standard[28]				
국가/기관	호주	자료 유형	지침·가이드라인	발표 일자	2024. 9. 5.

- 호주는 '24년 9월 '자발적 AI 안전표준(Voluntary AI Safety Standard)'을 발표하며 데이터 품질·모델 성능에 대한 배포 전·후 테스트, 수용 기준 설정 등을 통해 정확성을 포함한 AI 시스템의 성능과 위험을 지속적으로 평가·관리하는 방안 제시
  - **(데이터 품질: Guardrail 3)** AI 시스템의 데이터 품질 및 데이터 출처를 관리하기 위한 데이터 거버넌스 조치를 마련해야 함
  - **(배포 전·후 모델 성능평가: Guardrail 4)** AI 모델·시스템은 배포 전 테스트를 통해 모델 성능을 평가하고, 배포 후 시스템 행동 변화와 의도하지 않은 결과를 모니터링해야 함
  - **(수용기준: Guardrail 4)** AI 시스템의 잠재적 위해를 적절히 통제하기 위해 명확하고 측정가능한 수용 기준을 설정하고, 배포 전 해당 기준의 충족 여부를 테스트·평가하며, 시스템 변화에 따라 지속적으로 갱신해야 함

자료명	AI 편향 참조가이드 AI Bias Reference Guide[29]				
국가/기관	사우디아라비아	자료 유형	지침·가이드라인	발표 일자	2026. 1. 5.

- 사우디아라비아는 '26년 1월 'AI 편향 참조가이드(AI Bias Reference Guide)'를 통해 AI 시스템의 정확성과 공정성에 영향을 미치는 다양한 편향을 식별·완화하기 위한 가이드를 제시
  - **(편향 식별)** AI 시스템 정확성에 영향을 미칠 수 있는 100개 이상의 편향 유형 및 발생 원인 제시
  - **(영향 분석)** 편향이 AI 시스템의 의사결정 및 결과에 미치는 영향을 실제 사례를 통해 설명
  - **(편향 완화)** 각 편향 유형에 대한 완화전략을 제시하여 AI 시스템의 편향 영향을 줄이도록 지원

자료명	AI 윤리원칙 및 가이드라인 AI Ethics Principles & Guidelines[30]				
국가/기관	아랍에미리트	자료 유형	지침·가이드라인	발표 일자	2022. 12. 1.

- 아랍에미리트는 '22년 12월 'AI 윤리원칙 및 가이드라인(AI Ethics Principles & Guidelines)'에서 학습데이터의 정확성 유지, 시스템 오류의 추적·진단을 통해 데이터 정확성과 AI 시스템의 신뢰성을 높이기 위한 관리방안 제시
  - **(데이터 정확성)** 학습데이터의 출처·수집·처리 방법을 문서화하고 시간 경과에 따른 데이터 정확성 유지를 위한 조치를 마련하도록 권고
  - **(오류 관리)** AI 시스템의 실패를 추적·진단할 수 있도록 설계하고, AI로 인한 피해 발생 시 조사·시정 절차를 마련하도록 권고
  - **(모니터링)** AI 시스템의 취약성과 예상하지 못한 상황에서의 행동을 지속적으로 검증하고, 시스템이 목표·목적 및 의도된 적용 범위를 충족하는지 지속적으로 모니터링하도록 권고

□ 참고문헌

1) 이상욱. (2024). 안전하고 신뢰가능한 AI를 위한 국제 거버넌스와 AI 리터러시. Digital Economy View, 5, 14-25.
2) Federal Trade Commission. (2026, 7, 1). Federal Trade Commission's proposed policy statement concerning the suppression of accuracy in artificial intelligence systems.
3) Federal Trade Commission. (2026, 7, 1). Policy statement concerning the suppression of accuracy in artificial intelligence systems (Docket No. FTC-2026-0859) [Public comments docket].
4) Executive Order 14148, Initial rescissions of harmful executive orders and actions. (2025, 1, 20). Federal Register.
5) Executive Order 14179, Removing barriers to American leadership in artificial intelligence. (2025, 1, 23). Federal Register.
6) The White House. (2025, July). Winning the race: America's AI action plan.
7) Executive Order 14319, Preventing woke AI in the federal government. (2025, 7, 23). Federal Register.
8) Executive Order 14365, Ensuring a national policy framework for artificial intelligence. (2025, 12, 11). Federal Register.
9) M-26-04 Increasing public trust in artificial intelligence through unbiased AI principles. (2025, 12, 11). Office of Management and Budget.
10) Federal Trade Commission. (2026, 7, 1). Federal Trade Commission's proposed policy statement concerning the suppression of accuracy in artificial intelligence systems.
11) Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology.
12) Executive Order 14319, Preventing woke AI in the federal government. (2025, 7, 23). Federal Register.
13) M-26-04 Increasing public trust in artificial intelligence through unbiased AI principles. (2025, 12, 11). Office of Management and Budget.
14) Federal Trade Commission Act, 15 U.S.C. §§ 41-58 (1914).
15) Federal Trade Commission. (2026, 7, 1). Federal Trade Commission's proposed policy statement concerning the suppression of accuracy in artificial intelligence systems.
16) European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union.

17) Department for Science, Innovation and Technology, & Office for Artificial Intelligence. (2023, 8, 3). A pro-innovation approach to AI regulation (CP 815). GOV.UK.
18) Information Commissioner's Office. (2023, 3, 15). What do we need to know about accuracy and statistical accuracy? In Guidance on AI and data protection.
19) Information Commissioner's Office. (2023, 3, 15). Annex A: Fairness in the AI lifecycle. In Guidance on AI and data protection.
20) Government Digital Service. (2024, 12, 17). Algorithmic Transparency Recording Standard (ATRS) mandatory scope and exemptions policy. GOV.UK.
21) Treasury Board of Canada Secretariat. (2025, 6, 24). Directive on automated decision-making. Government of Canada.
22) Innovation, Science and Economic Development Canada. (2023, 9, 27). Voluntary code of conduct on the responsible development and management of advanced generative AI systems. Government of Canada.
23) AI Verify Foundation, & Infocomm Media Development Authority. (2024, 5, 30). Model AI governance framework for generative AI: Fostering a trusted ecosystem.
24) 国家互联网信息办公室. (2023, 7, 13). 生成式人工智能服务管理暂行办法.
25) 国家互联网信息办公室. (2026, 4, 10). 人工智能拟人化互动服务管理暂行办法.
26) 経済産業省. (2024, 4, 19). AI社会実装に向けた論点整理.
27) 數位發展部數位產業署. (2024, 4, 29). 人工智慧(AI)產品與系統評測參考指引(草案).
28) Department of Industry, Science and Resources. (2024, 9, 5). Voluntary AI safety standard: The 10 guardrails. Australian Government.
29) Saudi Data and Artificial Intelligence Authority [SDAIA]. (2026, 1, 5). AI bias reference guide. Government of Saudi Arabia.
30) Minister of State for Artificial Intelligence, Digital Economy and Remote Work Applications Office [AI Office]. (2022, 12, 1). AI ethics: Principles and guidelines. Government of the United Arab Emirates.

The LENS

The LENS는 주요국의 AI·디지털 정책 동향을 체계적으로 조사하고, 정기적 업데이트를 통해 정책 변화의 흐름을 지속적으로 추적·평가하는 정책 모니터링 보고서입니다.

| 작 성 |

- 한국지능정보사회진흥원 인공지능정책실 미래전략팀

  오 연 주 수석연구원 (053-230-1295, oyeonj@nia.or.kr)

| 기 획 |

- 한국지능정보사회진흥원 인공지능정책실

  이 용 진 실장　　정 지 선 팀장

  오 연 주 수석연구원

1. 본 보고서는 방송통신발전기금으로 수행한 과학기술정보통신부 정보통신·방송 연구개발사업(ICT진흥 및 혁신기반조성(정보화, R&D)사업)의 연구결과입니다.