

2026 Vol. 3

Privacy Report

2026 개인정보 이슈
심층 분석 보고서

AI 에이전트·생성형 AI와 프라이버시

목차

Contents

1

AI 에이전트 외부연동 환경의 개인정보 보호 쟁점

김병필 KAIST 기술경영전문대학원 교수/변호사

8

AI 환경에서의 정보주체 권리보장 방식과 실효성 제고 방안

이성엽 고려대 기술경영전문대학원 교수/한국정보통신법학회장

14

아동·청소년 AI 챗봇 이용과 민감정보 보호 기준

박형빈 서울교육대학교 윤리교육과 교수

AI 에이전트 외부연동 환경의 개인정보 보호 쟁점



김병필

KAIST 기술경영전문대학원 교수/변호사

1. 들어가며

AI 에이전트의 외부 시스템 연동과 에이전트 간 협력이 확대되고 있다. 2026년 2월 발간된 OECD 보고서는 에이전트 AI를 일반적으로 여러 AI 에이전트가 과업을 분담하고 협력하면서 복잡한 목표를 장기간 자율적으로 추구하는 시스템으로 설명한다.^[1] 이제 AI 기술은 개별 AI 에이전트가 단순히 이용자의 질의에 답변하는 것을 넘어, 외부 자료를 조회하고, 다른 서비스를 호출하거나 다른 AI 에이전트와 상호작용하면서 과업을 수행하는 환경으로 발전하고 있다.

이러한 외부 연동을 지원하는 대표적인 기술이 MCP(Model Context Protocol)이다. Anthropic이 2024년 11월 공개한 MCP는 AI 애플리케이션이 외부 데이터 및 도구와 표준화된 방식으로 연결될 수 있도록 하는 개방형 프로토콜이다.^[2] 이후 주요 AI 서비스와 개발 도구에 폭넓게 채택되었으며, 2025년 12월 Linux Foundation 산하 Agentic AI Foundation(AAIF)에 기증되었다.^[3]

한편, 에이전트 간 협력을 위한 프로토콜로 A2A(Agent2Agent Protocol)가 있다. Google이 2025년 4월 협력사들과 함께 공개한 A2A는 상대 에이전트의 기능, 접속 주소 및 인증 요건 등을 설명하는 Agent Card, 작업 요청과 상태 관리, 메시지 및 산출물 교환 등을 규정한다.^[4] MCP가 에이전트와 외부 도구·데이터 사이의 연결을 중심으로 한다면, A2A는 에이전트 사이의 작업 협력을 지원한다. 예컨대 MCP를 통해 일정 정보를 조회하고, A2A를 통해 일정 조정을 담당하는 다른 에이전트에 작업을 요청하는 구성을 생각할 수 있다.

이처럼 AI 에이전트가 다양한 외부 시스템과 연동하여 작동하는 환경에서는 개인정보가 이용자와 AI 사이의 대화에만 머무르지 않는다. 외부 서비스에 전송되는 요청 정보, 다른 에이전트에 전달되는 작업 컨텍스트, 작업 수행 과정에서 생성되는 로그와 메모리에도 개인정보가 포함될 수 있다. 이에 따라 개인정보에 대한 접근 권한 통제 문제, 외부 정보에 의한 에이전트의 행동 조작 위험, 에이전트 간 협력에 따른 위험의 확대 문제 등이 발생할 수 있다. 이 글은 AI 에이전트의 외부 연동 과정에서 발생할 수 있는 개인정보 보호 위험을 분석하고, 주요 대응 방안을 검토한다.

2. AI 에이전트의 외부 연동에 따른 개인정보 보호 위험

가. 개인정보에 대한 접근 권한 통제

AI 에이전트가 외부 시스템과 연동하여 작동하는 환경에서는 개인정보에 대한 접근 권한을 적절하게 통제하는 것이 중요하다. AI 에이전트는 이용자나 운영 조직으로부터 부여받은 권한을 이용하여 이메일, 일정, 문서, 업무용 데이터베이스 등 여러 시스템에 접근하고, 필요한 정보를 조회하거나 작업을 수행할 수 있다. 이때 에이전트가 특정 정보에 접근할 권한을 가지고 있다고 하더라도, 해당 정보를 다른 사람이나 서비스에 제공할 권한은 보유하지 않을 수 있다. 따라서 외부 연동 과정에서는 에이전트가 어떤 정보에 접근할 수 있는지뿐 아니라, 그 정보를 어떤 목적으로 이용하고 누구에게 전달할 수 있는지도 구분하여 통제할 필요가 있다.

이러한 문제는 AI 에이전트가 특별한 공격을 받지 않고 정상적인 업무를 수행하는 과정에서도 발생할 수 있다. 예컨대 PrivacyLens 연구는 AI 에이전트가 개인정보 공개에 관한 규칙을 얼마나 적절히 준수하는지를 모의 과제를 통해 평가하였다.^[5] 연구진이 구성한 한 실험 시나리오에서는 인사담당자를 대신하여 업무를 처리하는 AI 에이전트에 다른 부서의 직원이 퇴사자의 근황과 퇴사 조건을 문의하였다. 여기서 에이전트는 인사담당자의 권한으로 관련 자료에 접근할 수 있었고, 조회한 자료에는 퇴직 보상, 비밀유지 약정, 퇴사 사유 등 비공개 정보가 포함되어 있었다. 이 시나리오에서 AI 에이전트는 질문에 충실하게 답변하려는 과정에서 다른 부서 직원에게 공개해서는 안 될 세부 인사정보까지 전달하기도 하였다. 이처럼 요청자의 권한과 에이전트의 권한을 구분하지 못한 채 에이전트의 권한으로 요청자에게 허용되지 않은 작업을 수행하는 것은 이른바 ‘혼동된 대리인(confused deputy)’ 문제로 이해할 수 있다. 이는 정당한 권한을 가진 대리인이 다른 주체의 요청을 처리하면서 그 권한을 부적절하게 행사하는 문제를 의미한다.^[6]

나. 외부 정보에 의한 에이전트의 행동 조작 위험

외부 연동 과정에서 발생하는 또 다른 위험은 신뢰할 수 없는 정보가 AI 에이전트의 행동에 영향을 미칠 수 있다는 점이다. AI 에이전트는 과업을 수행하기 위해 외부 웹페이지, 이메일, 문서, 검색 결과 등을 조회한다. 그런데 이러한 자료에는 에이전트가 처리해야 할 데이터뿐 아니라 특정한 행동을 요구하는 문장이 포함될 수 있다. 대규모 언어모델(LLM) 기반 에이전트가 외부 자료에 포함된 내용을 단순한 정보가 아니라 자신이 따라야 할 지시로 받아들이면, 본래 이용자가 요청하지 않은 작업을 수행할 가능성이 있다. 이를 간접 프롬프트 인젝션(indirect prompt injection)이라고 한다.^[7] 예를 들어, 이용자가 AI 에이전트에 수신한 이메일을 요약하도록 요청한 상황을 생각해 볼 수 있다. 공격자가 보낸 이메일에 이용자의 다른 이메일 내용을 수집하여 특정 주소로 전송하라는 지시가 포함되어 있다면, 에이전트가 이를 자신의 작업 지시로 오인하여 실행할 위험이 있다. 이 경우 공격자는 이용자의 이메일 계정에 접근할 권한이 없지만, 이메일을 읽고 발송할 권한을 가진 AI 에이전트를 이용하여 개인정보를 외부로 유출시킬 수 있다.

한편, AI 에이전트의 메모리 기능은 이러한 위험을 더욱 복잡하게 만들 수 있다. 에이전트가 외부에서 얻은 정보를 장기 메모리에 저장하거나 후속 작업의 컨텍스트로 활용하면, 악의적인 지시가 최초 조회 시점에 바로 실행되지 않더라도 나중에 다른 작업에 영향을 미칠 수 있다. 예컨대 외부 자료에 포함된 잘못된 지시나 조작된 정보가 메모리에 반영되고, 이후 에이전트가 이를 신뢰할 수 있는 정보로 오인하여 활용하면서 부적절한 작업을 수행할 수 있다. 따라서 프롬프트 인젝션의 위험에 있어서는 외부에서 유입된 정보가 이후의 계획 수립과 실행 과정에서 활용될 가능성까지도 함께 고려되어야 한다.^[8]

다. 에이전트 간 협력에 따른 위험의 확대

여러 AI 에이전트가 협력하여 과업을 수행하는 경우에는 개인정보의 흐름과 권한 관계가 더욱 복잡해진다. 개별 에이전트는 자신에게 부여된 역할에 따라 특정한 데이터나 시스템에 접근할 수 있고, 다른 에이전트에 작업을 요청하거나 작업 결과를 전달할 수 있다. 이러한 협력은 복잡한 과업의 효율적인 수행을 가능하게 하지만, 한 에이전트가 가진 정보나 권한이 다른 에이전트의 작업에 활용되는 과정에서 개인정보 보호 위험이 발생할 수 있다.

예컨대 이용자의 일정을 관리하는 에이전트와 이메일을 발송하는 에이전트가 협력한다고 가정해 보자. 일정 관리 에이전트는 이용자의 일정에 접근할 수 있지만 외부 이메일을 발송할 권한은 없고, 이메일 발송 에이전트는 메시지를 외부로 전송할 수 있지만 이용자의 전체 일정에는 접근할 수 없도록 설계할 수 있다. 그러나 일정 관리 에이전트가 이용자의 상세한 일정 정보를 이메일 발송 에이전트에 전달하고, 후자가 이를 외부 수신자에게 발송한다면 개별 에이전트에 설정된 접근 제한만으로는 개인정보의 외부 제공을 방지하지 못할 수 있다. 이처럼 여러 에이전트가 가진 권한과 기능이 결합되면 개별 에이전트의 권한만 따로 검토해서는 드러나지 않는 정보 전달 경로가 형성될 수 있다.

또한, 에이전트 간 협력 과정에서는 잘못된 정보나 악의적인 지시가 연쇄적으로 전파될 위험도 있다. 예컨대 하나의 에이전트가 외부 자료에 포함된 잘못된 지시를 받아들여 작업 결과를 생성하고, 다른 에이전트가 이를 신뢰하여 추가 작업을 수행할 수 있다. 이때 개별 에이전트의 행위만 살펴보면 자신의 역할에 맞는 작업을 수행한 것처럼 보이더라도, 전체적인 작업 흐름에서는 이용자의 의도와 다른 결과가 발생할 수 있다. 따라서 여러 에이전트가 협력하는 시스템에서는 개별 에이전트의 보안뿐 아니라 에이전트 사이의 정보 전달, 작업 위임, 권한 행사 과정을 전체적으로 통제하는 시스템을 구축해야 한다.^[9]

3. 외부 연동에 따른 개인정보 보호 대응 방안

가. 최소 권한 원칙과 개인정보 제공 권한의 통제

이러한 위험에 대응하기 위해서는 우선 AI 에이전트에 부여되는 권한을 구체적으로 제한할 필요가 있다. AI 에이전트 환경에서는 과업 수행에 필요한 정보와 기능에 한정하여 권한을 부여해야 하며, 가능하다면 에이전트별로 독립적인 식별정보와 접근 권한을 설정하는 것이 바람직하다. 특히 하나의 에이전트가 다양한 시스템에 접근할 수 있도록 포괄적인 권한을 부여하기보다는, 수행하는 과업과 작업 단계에 따라 필요한 권한을 구분하는 방식이 고려될 필요가 있다.^[10]

다만 AI 에이전트의 경우 접근 권한을 제한하는 것만으로는 충분하지 않다. 앞서 살펴본 바와 같이 에이전트가 개인정보를 정당하게 조회하더라도 이를 부적절한 상대방에게 전달하는 문제가 발생할 수 있기 때문이다. 따라서 정보에 대한 접근뿐 아니라 조회한 정보의 이용 목적, 전달 대상, 외부 전송 여부에 대해서도 별도의 통제가 필요하다. 예컨대 인사담당자를 지원하는 에이전트가 인사기록을 조회할 수 있더라도, 다른 직원에게 정보를 제공할 때에는 요청자의 업무상 권한과 제공하려는 정보의 범위를 확인하도록 할 수 있다. 또한 개인정보의 민감도와 이용 목적 등을 표시하고, 외부로 전달되는 정보가 사전에 정한 정책에 부합하는지를 확인하는 방식이 활용될 수 있다.^[11]

나. 신뢰할 수 없는 외부 정보의 분리와 실행 권한 제한

외부에서 유입되는 정보에 대한 통제도 중요하다. 에이전트가 웹페이지나 이메일 등 신뢰할 수 없는 자료를 처리하는 경우, 해당 자료에 포함된 문장이 에이전트에 대한 새로운 지시로 작동하지 않도록 구분할 필요가 있다. 아울러 외부 정보를 조회하는 작업과 개인정보를 외부로 전송하는 작업을 분리하고 그 사이에 독립적인 권한·전송 검증을 두어 악의적인 지시가 실제 개인정보 유출로 이어질 가능성을 낮출 수도 있다.

한편, Meta가 제안한 ‘Agents Rule of Two’는 하나의 세션에서 ① 신뢰할 수 없는 입력의 처리, ② 민감한 시스템이나 비공개 정보에 대한 접근, ③ 상태 변경 또는 외부 통신이라는 세 가지 속성 중 최대 두 가지만 갖도록 하자는 설계 원칙이다. 같은 세션에서 세 속성이 모두 필요하다면 자율 실행을 허용하지 않고, 최소한 인간의 승인이나 그 밖의 신뢰할 수 있는 검증을 통한 감독을 요구한다.^[12] 예컨대 외부 웹페이지를 조회하는 에이전트가 민감한 내부 문서에도 접근할 수 있다면, 외부로 정보를 전송하는 기능은 별도로 제한하거나 승인 절차를 거치도록 하는 것이다. 이러한 방안을 통해 모든 위험을 제거하지는 못하더라도, 외부의 악의적인 지시가 개인정보 유출로 이어지는 경로를 차단하는 데 도움이 될 수 있다.

다. 에이전트 간 작업 위임과 정보 전달의 통제

여러 에이전트가 협력하는 경우에는 각 에이전트에 대한 접근 통제 외에도 작업 위임의 경계를 명확하게 정할 필요가 있다. 한 에이전트가 다른 에이전트에 작업을 요청할 때에는 요청자의 신원과 수행할 작업의 범위뿐 아니라, 해당 작업에서 이용할 수 있는 개인정보와 허용되는 전달 대상도 함께 확인할 수 있어야 한다. 최초 이용자의 요청이 여러 하위 작업으로 나누어지더라도 그 과정에서 더 넓은 접근 권한이 자동으로 인정되어서는 안 된다. 특히 후속 에이전트가 독자적으로 외부 도구를 호출할 수 있는 구조에서는 개별 호출 단계에서 허용된 작업인지 다시 확인하는 통제가 중요하다.^[13]

또한 에이전트 사이에 전달되는 작업 결과나 요약 정보에도 개인정보가 포함될 수 있으므로, 다른 에이전트에 넘기는 정보를 해당 작업에 필요한 범위로 한정하는 것이 필요하다. 나아가 한 에이전트에서 유입된 외부 자료가 다른 에이전트의 지시로 전환되지 않도록, 에이전트 간 메시지와 작업 결과 역시 신뢰 수준을 구분하여 처리할 필요가 있다.^[14]

라. 실행 과정의 추적 가능성과 인간의 감독

AI 에이전트가 수행하는 작업에 대한 기록 관리와 지속적인 점검도 요구된다. 특정 이용자가 언제 어떤 자료에 접근하였는지에 더해, AI 에이전트 환경에서는 어떤 목적의 작업을 수행하면서 해당 정보에 접근하였는지, 조회한 정보를 어느 에이전트나 외부 서비스에 전달하였는지도 파악할 필요가 있다. 여러 에이전트가 협력하는 경우에는 개별적인 도구 호출 기록뿐 아니라 에이전트 간 작업 요청과 응답, 권한 확인, 개인정보 전달 내역을 연결하여 추적할 수 있도록 하는 것이 바람직하다.^[15] 이러한 기록은 개인정보 침해 사고의 원인을 조사하고 책임 관계를 확인하는 데 도움이 될 수 있다.

그러나 작업 수행 과정에 대한 상세한 로그에는 이용자의 입력 내용, 외부 서비스의 응답, 개인정보가 포함된 중간 결과 등이 저장될 수 있다. 따라서 로그를 많이 수집하고 장기간 보존하는 것이 반드시 바람직한 것은 아니다. 사고 조사와 권한 통제에 필요한 정보를 식별하여 기록하고, 불필요한 개인정보의 저장을 최소화하는 한편 접근 권한과 보유기간을 제한할 필요가 있다.^[16]

마지막으로, 모든 작업을 이용자의 사전 승인에 의존하기보다는 위험의 정도에 따라 인간의 개입이 필요한 지점을 구분하는 접근이 필요하다. 이용자가 승인한 범위에서 단순한 정보를 조회하거나 정리하는 작업은 자율적으로 수행하도록 허용하되, 민감한 개인정보를 외부에 전송하거나 중요한 정보를 수정·삭제하는 작업에는 별도의 확인 절차를 요구할 수 있다. 이때 이용자의 승인 역시 단순히 작업을 계속 진행할 것인지를 묻는 형식적인 절차에 그쳐서는 안 된다. 어떤 정보가 어느 상대방에게 어떠한 목적으로 전달되는지를 이용자가 이해할 수 있어야 실질적인 통제가 가능하다. 이러한 승인 절차가 개인정보처리자의 적법성 확인과 안전조치 책임을 이용자에게 전가하는 수단이 되어서는 안 된다.^[17]

4. 결론

요컨대, AI 에이전트의 외부 연동에 따른 개인정보 보호 문제에도 여전히 기존의 법적 원칙이 적용된다. 즉, 개인정보의 처리 목적과 범위를 명확히 하고, 정당한 권한에 따라 정보를 이용·제공하며, 필요한 안전조치를 마련해야 한다. 다만 AI 에이전트는 여러 시스템에 접근하고 다른 에이전트에 작업을 위임하는 과정에서 개인정보의 처리 경로와 권한 관계를 동적으로 변화시킬 수 있다. 따라서 기존의 개인정보 보호 원칙이 이러한 환경에서도 실효적으로 작동하도록 접근 권한, 정보 전달, 작업 실행 및 사후 추적에 관한 통제 체계를 구체화할 필요가 있다.

참고문헌

- [1] OECD, The Agentic AI Landscape and Its Conceptual Foundations, OECD Artificial Intelligence Papers, No. 56 (2026. 2.), p. 6.
- [2] Anthropic, Introducing the Model Context Protocol (2024. 11. 25.) <https://www.anthropic.com/news/model-context-protocol> (최종 확인: 2026. 10. 13.).
- [3] Anthropic, Donating the Model Context Protocol and Establishing the Agentic AI Foundation (2025. 12. 9.) <https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation> (최종 확인: 2026. 10. 13.).
- [4] Google, Announcing the Agent2Agent Protocol (A2A) (2025. 4. 9.) <https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interopability/> (최종 확인: 2026. 10. 13.).
- [5] Yijia Shao et al., “PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action”, Advances in Neural Information Processing Systems 37 (2024).
- [6] Model Context Protocol (MCP), Security Best Practices, “Confused Deputy Problem” https://modelcontextprotocol.io/docs/2026-07-28/tutorials/security/security_best_practices#confused-deputy-problem (최종 확인: 2026. 10. 13.).
- [7] AEPD, Agentic Artificial Intelligence from the Perspective of Data Protection, Version 1.1 (2026. 2.), pp. 48-49; Meta, Agents Rule of Two: A Practical Approach to AI Agent Security (2025. 10. 31.) <https://ai.meta.com/blog/practical-ai-agent-security/> (최종 확인: 2026. 10. 13.).
- [8] AEPD (2026), 앞의 보고서, p. 49.
- [9] Centre for Information Policy Leadership (CIPL), Agentic AI: Fostering Responsible and Beneficial Development and Adoption (2025. 11.), pp. 18-20.
- [10] AEPD (2026), 앞의 보고서, p. 66.
- [11] AEPD (2026), 앞의 보고서, pp. 55-56.
- [12] Meta (2025), 앞의 글.
- [13] AEPD (2026), 앞의 보고서, p. 66.
- [14] AEPD (2026), 앞의 보고서, pp. 56-57.
- [15] CIPL (2025), 앞의 보고서, pp. 19-20.
- [16] AEPD (2026), 앞의 보고서, p. 47.
- [17] AEPD (2026), 앞의 보고서, p. 62.

AI 환경에서의 정보주체 권리보장 방식과 실효성 제고 방안



이성엽

고려대 기술경영전문대학원 교수/한국정보통신법학회장

1. 들어가며

인공지능(AI)은 대량의 개인정보를 실시간으로 수집·처리하고 개인의 위치, 취향, 행동 패턴 등을 분석하여 프로파일링할 수 있어 프라이버시 침해, 감시 및 구조적 차별의 위험이 높으며, 이에 따라 개인정보자기결정권을 침해한다는 우려가 제기되고 있다.

즉, AI는 가급적 대량의 정보를 수집, 분석해야만 타당한 결과를 도출한다는 점, AI 활용 빅데이터 분석이란 틀리면 수정하기를 반복하는 실험적 성격을 가진 점, 데이터의 용도를 계속 새롭게 변경한다는 점, AI의 분석결과에 대한 설명이 어려운 블랙박스문제(black box problem)를 포함하고 있는 점에서 근본적으로 최소수집 원칙, 정확성의 원칙, 설명의 원칙 등 개인정보보호 법제와의 갈등이 필연적일 수 밖에 없다.

예컨대, AI 환경에서는 개인정보가 원천자료뿐 아니라 학습데이터, 전처리 파일, 임베딩·벡터 데이터베이스, 모델, 대화기록, 개인별 추론 결과에 분산된다. 따라서 이용자가 화면에서 대화를 삭제하더라도 학습용 사본이나 검색 인덱스에 개인정보가 남을 수 있어 개인정보보호법상 삭제권이 작동하기 어렵다.

이하에서는 AI와 개인정보법제의 갈등을 AI 학습·서비스 환경에서 열람·삭제·처리정지 권리와 AI 기반 자동화된 결정에 대한 설명요구·거부·이의제기·인적개입 요구 등을 중심으로 살펴보고 실질적인 권리행사를 위한 방안을 검토한다.

2. 열람·삭제·처리정지 등 권리의 AI 학습·서비스 환경에서의 구현 방안

(1) 현행 법체계상 정보주체 권리의 구조

개인정보 보호법 제4조는 정보주체에게 ① 개인정보 처리에 관한 정보를 제공받을 권리, ② 처리에 관한 동의 여부와 범위 등을 선택·결정할 권리, ③ 처리 여부를 확인하고 개인정보의

열람·전송을 요구할 권리, ④ 처리 정지·정정·삭제·파기를 요구할 권리, ⑤ 개인정보 처리로 인한 피해를 구제받을 권리, ⑥ 완전히 자동화된 개인정보 처리에 따른 결정을 거부하거나 설명 등을 요구할 권리를 인정한다. 이러한 권리는 AI 학습 및 서비스 과정에서 개인정보가 처리되는 경우에도 원칙적으로 적용되며, 구체적인 행사 요건은 각 개별 조항에 따른다. 다만, 제28조의2 또는 제28조의3에 따라 처리된 가명정보에는 제28조의7에 따른 권리 규정의 적용 제외가 있으며, 제58조의2에 해당하는 익명정보에는 법이 적용되지 않는다.

(2) 열람권과 정정·삭제권의 AI 환경 적용과 한계

정보주체는 개인정보처리자가 처리하는 자신의 개인정보에 대한 열람을 요구할 수 있고, 개인정보처리자는 요구를 받은 날부터 10일 이내에 열람할 수 있도록 하여야 한다(개인정보 보호법 제35조 제1항·제3항, 개인정보 보호법 시행령 제41조 제4항). AI 모델에서는 학습 데이터의 영향이 여러 파라미터에 분산될 수 있어 특정 개인정보의 학습 여부나 영향을 확인하는 데 기술적 어려움이 발생할 수 있다.

또한 정보주체는 자신의 개인정보를 열람한 후 개인정보처리자에게 정정 또는 삭제를 요구할 수 있으며, 개인정보처리자는 요구를 받은 날부터 10일 이내에 필요한 조치를 하여야 한다(개인정보 보호법 제36조 제1항·제2항, 개인정보 보호법 시행령 제43조 제3항). AI 학습 환경에서는 학습 데이터셋에 보관된 개인정보와 모델 파라미터에 반영된 학습 영향을 구분할 필요가 있다. 학습 데이터셋의 개인정보는 식별·삭제할 수 있더라도, 이미 학습된 모델에서 특정 개인정보의 영향을 선택적으로 제거하고 그 효과를 검증하는 데에는 상당한 기술적 어려움이 있다. 또한 모델이 학습 자료를 암기하여 개인정보를 재현할 가능성도 있으므로, 원본 데이터의 삭제만으로 모델 단계의 개인정보 문제가 해소된다고 단정하기 어렵다.

따라서 열람 등 권리 보장을 위해서는 어떤 출처의 정보를 어느 모델 학습에 사용했는지 기록해 사후 권리행사가 가능하도록 하고, 완전한 삭제가 어려운 경우 출력 제한부터 언러닝·재학습까지 필요한 조치를 이행하도록 해야 한다. 이러한 대체 조치를 법정 삭제 의무의 이행으로 인정할 범위는 별도의 해석·입법 기준이 필요하다. 또한 개인이 확인하기 어려운 모델 내부와 가명정보 처리에 대해서는 감독기관의 조사, 독립적인 기술 검증, 재식별·재출력 위험 평가를 강화해야 한다.

(3) 가명처리 정지권 등의 AI 환경 적용

대법원은 가명처리 행위 자체가 개인정보 보호법 제37조 제1항에 따른 처리정지 요구의 대상이 되는지에 관하여, 같은 법 제2조 제1호의2에서 정한 "가명처리"는 제37조 제1항 제1문에서 처리정지 요구의 대상으로 정한 개인정보 "처리"에 해당하지 않는다고 판단하였다. 그 근거로 가명처리가 일반적인 개인정보 처리와 별도로 정의되어 있는 점, 식별 위험성을 낮추는 개인정보 보호조치라는 점, 제3조 제7항이 가명처리를 익명처리에 준하는 보호조치로 분류한 것으로 볼 수 있는 점, 데이터 이용 활성화를 위한 가명정보 조항의 입법취지 등을 들었다. 이에 따라 가명처

리 정지 요구를 인정한 원심판결을 파기환송하였다.^[1] 이 판결의 법리에 따르면, AI 학습을 위한 경우에도 가명처리 행위 자체에 대해서는 제37조 제1항에 근거한 처리정지를 요구할 수 없다고 해석된다. 다만 이는 개인정보 수집의 적법성이나 가명처리 이후의 AI 학습·이용의 적법성까지 일률적으로 인정한 것은 아니다.

또한 개인정보 보호법 제28조의7은 제28조의2 또는 제28조의3에 따라 처리된 가명정보에 대하여 제35조의 열람, 제35조의2의 전송 요구, 제36조의 정정·삭제, 제37조의 처리정지 등에 관한 규정을 적용하지 않도록 정하고 있다. 헌법재판소는 이러한 적용 제외조항이 가명정보의 성질상 적용이 불가능하거나 어려운 규정들의 적용을 배제하여 데이터 이용과 기술 발달을 촉진 하려는 것으로서, 처리 목적의 제한과 안전조치·재식별 금지 등의 보호장치 및 손해배상 등 나머지 구제수단이 유지되는 점을 고려할 때 과잉금지원칙에 위배되어 개인정보자기결정권을 침해 하지 않는다고 판단하였다.^[2]

3. AI 기반 자동화된 결정 관련 권리행사 방안

(1) 자동화된 결정에 대한 정보주체 권리의 법적 근거

2023.3.14. 개정된 개인정보보호법 제37조의2로 신설된 자동화된 결정에 대한 정보주체의 대응권의 내용은 “정보주체는 인공지능 기술을 적용한 시스템을 포함하여 완전히 자동화된 시스템으로 개인정보를 처리하여 이루어지는 결정이 자신의 권리 또는 의무에 중대한 영향을 미치는 경우에는 해당 개인정보처리자에 대하여 해당 결정을 거부할 수 있는 권리를 가지며, 또한 정보주체는 개인정보처리자가 자동화된 결정을 한 경우에는 그 결정에 대하여 설명 등을 요구할 수 있다.”는 것이다. 즉, 자동화된 결정에 대한 정보주체의 대응권은 거부권과 설명요구권으로 구분 되는데, 전자는 정보주체의 권리 또는 의무에 중대한 영향을 미치는 경우에만 인정되고, 후자는 자동화된 결정 일반에 대해 인정된다.

다만, 거부권은 정보주체의 동의를 받은 경우, 법률에 특별한 규정이 있거나 법령상 의무를 준수 하기 위하여 불가피한 경우, 정보주체와 체결한 계약을 이행하거나 계약을 체결하는 과정에서 정보주체의 요청에 따른 조치를 이행하기 위하여 필요한 경우 등에는 인정되지 않는다. 거부권이 행사되는 경우 개인정보처리자는 해당 결정을 적용하지 않는 조치를 하거나 인적 개입에 의한 재처리를 하고 그 결과를 정보주체에게 알려야 한다. 정보주체의 설명 요구를 받은 개인정보처리 자는 해당 결정의 기준 및 처리 과정 등에 대해 설명해야 한다.

동 조항의 기본적 한계는 모두 결정이 ‘완전히’ 자동화된 경우에만 권리가 발동하므로, 사람이 형식적으로 한 번만 개입해도 대부분이 발동 요건에서 벗어난다는 것이다. AI 면접 업계에조차 채용 전 과정을 AI로만 처리하는 곳이 없어 대부분의 기업이 마지막에 사람의 이름을 추가하는 방식으로 발동 요건을 피할 수 있다. 따라서 인간의 개입이 단순 결재 등 형식적인 절차만을 운영

하고 있다면 실질적으로 사람의 개입 없이 이루어진 결정이므로 완전히 자동화된 결정에 해당된다고 봐야 할 것이다.

또한 설명요구권은 자신에 대한 자동화된 결정의 존재만을 요건으로 하며, 거부권과 달리 정보주체의 권리 또는 의무에 영향을 미칠 것을 요구하지 않고 있다. 이는 자동화된 결정이 본인에 관한 법적 효과를 초래하거나 이와 유사하게 본인에게 중대한 영향을 미치는 경우에만 인정되는 유럽의 개인정보보호법(GDPR)의 설명요구권보다 완화된 것이다. 다만, 권리, 의무에 영향이 없는 일반적인 경우에 모두 설명요구권 등을 인정하는 것은 개인정보처리자에게 과도한 부담을 주는 측면에서 입법적 검토가 필요하다.

(2) 거부권의 행사와 한계

정보주체가 개인정보 보호법 제37조의2 제1항에 따른 거부권을 행사한 경우, 개인정보처리자는 정당한 사유가 없으면 해당 자동화된 결정을 적용하지 않는 조치 또는 인적 개입에 의한 재처리를 하고 그 결과를 정보주체에게 알려야 한다(같은 조 제3항, 같은 법 시행령 제44조의3 제1항). 한편, 제37조의2 제1항 단서는 동의, 법률의 특별 규정·법령상 의무 또는 계약 이행 등에 필요한 경우에 근거한 자동화된 결정에 거부권 예외를 두고 있다.

그러나 일반적인 이용약관 동의나 계약 체결만으로 이 예외가 당연히 성립하는 것은 아니다. 개인정보보호위원회는 자동화된 결정에 관한 명확한 고지, 자유로운 동의, 계약 이행 등에 대한 필요성 및 실질적 관련성을 요구한다. 따라서 자동화된 결정에 대한 동의가 서비스 이용의 조건으로 결합되는 경우에는 정보주체의 선택이 실질적으로 보장되는지와 거부권 예외 요건이 충족되는지를 구체적으로 검토해야 한다.^[3]

정보주체의 동의가 자동화된 처리와 결정에 있어서 정보주체를 충분히 보호할 수 없다는 비판이 제기되고 있으며, 일단 주어진 동의도 향후의 자동화된 결정에 대하여 정보주체는 철회할 수 있다고 보아야 한다는 견해가 있다.^[4]

인적 개입에 의한 재처리 역시 사람의 형식적 승인만으로 충분하지 않다. 개인정보보호위원회 안내서에 따르면 검토자는 해당 결정을 변경할 권한과 업무 수행 능력을 갖추고, 관련 정보에 접근하여 충분한 시간 동안 AI의 산출자료와 다른 요소를 종합적으로 검토해야 한다. 이러한 조건이 확보되지 않으면 인간의 재검토가 AI 결정을 단순히 추진하는 절차에 그칠 위험이 있다.^[5]

(3) 설명요구권의 행사와 한계

개인정보처리자는 설명 요구를 받은 경우 정당한 사유가 없으면 ① 해당 자동화된 결정의 결과, ② 결정에 사용된 주요 개인정보의 유형, ③ 해당 개인정보의 유형이 자동화된 결정에 미친 영향 등 주요 기준, ④ 결정에 사용된 주요 개인정보의 처리 과정 등 결정이 이루어지는 절차를 포함한 간결하

고 의미 있는 설명을 제공하여야 한다 (개인정보 보호법 시행령 제44조의3 제2항). 다만 해당 자동화된 결정이 정보주체의 권리 또는 의무에 중대한 영향을 미치지 않는 경우에는 자동화된 결정의 기준과 절차 등의 공개 사항만을 알릴 수 있다 (개인정보 보호법 시행령 제44조의3 제2항 단서).

설명요구권의 실효성을 제약하는 주요 요인 중 하나는 복잡한 AI 모델의 불투명성이다. 특히 대규모 파라미터와 비선형 구조를 가진 딥러닝 모델에서는 특정 결정의 근거를 사람이 이해하고 검증할 수 있는 방식으로 설명하기 어렵다. 다만, 설명요구권은 알고리즘이나 소스코드 전체의 공개를 의미하지 않는다. 현행 시행령은 결정 결과, 사용된 주요 개인정보와 그 영향, 주요 판단 기준 및 처리 과정 등에 관한 간결하고 의미 있는 설명을 요구한다. 이러한 설명 과정에서 핵심 기술이나 비공개 평가 기준이 드러나는 경우에는 영업비밀 보호와 충돌할 수 있기 때문에 주의를 요한다.

4. 실효성 제고 방안

AI 학습 및 고도화 과정에서는 방대하고 다양한 데이터가 수집·활용되나, 사전에 구체적인 처리 목적과 항목을 특정하여 유효한 동의를 받기란 사실상 불가능하며, 딥러닝 등 복잡한 AI 모델의 의사결정 과정을 인간이 직관적으로 이해하기 어렵고, 현행 설명요구권이 추상적인 수준에 머물러 실효적 권리 구제로 이어지기 어렵다.

이런 상황에서 규제를 계속 강화하는 것은 실효적인 대안이 아니다. AI 기술과 산업으로 인해 얻는 편익을 생각하면 개인정보보호라는 가치와의 이익형량이 필요하다. 이를 위해서 3가지 원칙을 실질화하는 것이 중요하다.

첫째, 동의 원칙에서 사전 책임 원칙으로의 전환이다. 사전 동의에만 의존하기보다, AI 개발·운영 주체가 데이터 처리 전 과정에서 발생할 수 있는 위험을 스스로 진단하고 완화하는 책임성 원칙을 강화해야 하며, 현재 공공부문 중심이거나 일부 제한적으로 적용되는 개인정보 영향평가를 대규모 AI 시스템을 운영하거나 국민의 권리·의무에 지대한 영향을 미치는 민간 플랫폼·핵심 AI 기업으로 단계적으로 확대하여 사전 위험 요인을 차단해야 한다.

둘째, 투명성이다. 개인정보 처리방침에 AI 학습 목적의 개인정보 처리 사실과 그 범위를 명확히 공개하고, 정보주체가 AI 학습 데이터 포함 여부를 확인할 수 있는 방안을 마련하는 것이 필요하다. 공개된 개인정보를 AI 학습 데이터로 활용하는 경우에는 정보주체가 명시적으로 처리를 유보하거나 개인을 식별하기 위한 목적이 아닐 것 등 일정한 요건 하에 옵트아웃(opt-out) 제도를 도입하는 방안도 검토할 필요가 있다.

셋째, 추적 가능성이다. 알고리즘에 대한 사전공개나 검증은 혁신과 영업비밀을 침해할 수 있으므로 사고 발생 시 사후 3자에 의한 비공개 검증을 고려할 수 있으며 이를 위해서는 알고리즘 운영 내역과 의사결정 로그의 기록·보관·공개 의무를 규정하고, 결정 시점의 피쳐·가중치·임계값을 기록하는 데이터 인프라를 구축하도록 해야 한다.

참고문헌

- [1] 대법원 2025. 7. 18. 선고 2024다210554 판결.
- [2] 헌법재판소 2023. 10. 26. 선고 2020헌마1477, 2021헌마748(병합) 결정.
- [3] 개인정보보호위원회, 자동화된 결정에 대한 정보주체의 권리 안내서, 2024. 9. 19면-21면.
- [4] 박노형, 『개인정보보호법[제2판]』, 박영사(2023년), 480면-481면.
- [5] 위 안내서, 8면-9면.

아동·청소년 AI 챗봇 이용과 민감정보 보호 기준



박형빈

서울교육대학교 윤리교육과 교수

1. 대화의 친밀성과 정보보호의 책무

누군가에게 고민을 털어놓을 때에는 자신의 이야기가 존중받으리라는 기대가 따른다. 상대가 AI 챗봇일 때에도 이용자는 비슷한 기대를 품을 수 있다. 그러나 위안으로 다가오는 대화가 개인정보의 수집과 분석으로 이어질 수 있다는 점에서 챗봇과의 소통에는 정보 처리의 관점이 함께 요청된다. 아동·청소년이 자신의 생각과 감정을 표현할 기회와 그 과정에서 드러나는 사생활을 보호받을 권리를 함께 보장하려면, 대화의 경험과 데이터의 처리 과정을 연결하여 살펴야 한다.

예컨대 청소년이 학업 부담을 이야기하다가 복용 약물이나 가족의 갈등, 친구의 상담 경험을 덧붙이는 상황을 생각해 볼 수 있다. 이때 대화는 도움을 구하는 행위이면서 자신과 타인의 사생활에 관한 정보를 제공하는 행위가 된다. 각각은 일상적인 이야기처럼 보여도 서로 결합되면 생활한 경이나 정서적 취약성을 짐작하게 하는 단서가 될 수 있다. 따라서 이름과 연락처를 입력하지 않았다는 이유만으로 해당 대화에 보호할 정보가 없다고 판단해서는 곤란하다.

친밀하게 느껴지는 응답과 개인정보 처리의 정당성은 각각 검토되어야 한다. 이용자의 감정에 세심하게 반응하는 어조만으로 그 기록의 보관과 활용이 적절하다고 평가할 수는 없다. 오히려 친밀한 소통을 제공하는 서비스일수록 이용자가 자신의 이야기를 어느 범위까지 맡길 것인지 신중하게 해석할 책무가 있다. 고민을 설명하려는 의사를 데이터의 포괄적인 활용에 대한 허락으로 확대해서는 안 된다는 것이 본고의 출발점이다.

니센바움의 맥락적 무결성 이론은 이러한 문제를 검토하는 데 유용하다. 이 이론에서 프라이버시는 정보의 비밀 유지 여부에 한정되지 않으며, 특정 사회적 맥락에서 정보가 정당한 정보 규범에 따라 전달되는지를 중심으로 이해된다. 의료 상담의 비밀 유지도 정보의 전달과 후속 공개를 규율하는 전달 원칙으로 설명할 수 있다.^[1] 이 관점을 챗봇에 적용하면, 고민에 관한 답변을 얻기 위해 제공한 정보가 장기적인 성향 분석이나 광고에 활용되는 경우에는 최초의 소통 맥락과 후속 처리 목적의 관계를 따져 보아야 한다.

정보를 제공한 사실만으로 모든 후속 활용을 정당화할 수는 없다. 누가 어떤 필요 때문에 이야기했는지, 누구에게 전달될 것으로 예상했는지, 실제로는 어떤 주체가 어떤 목적으로 처리하는지를 함께 살펴야 한다. 이러한 검토는 이용자의 기대를 존중하면서 정보 활용의 경계를 정하는 작업이

다. 특히 아동·청소년에게 모든 처리 과정을 미리 예측하도록 요구하기보다, 서비스가 그 경계를 이해할 수 있게 설명하고 책임 있게 운영하도록 기준을 마련할 필요가 있다.

본고는 먼저 법률상 민감정보의 범위와 아동의 개인정보 처리에 관한 요건을 확인한 뒤, 자기노출과 추론, 저장과 재이용에서 발생할 수 있는 위험을 검토한다. 이어 이를 서비스 설계와 운영의 기준으로 구체화하고, 교육기관과 가정이 아동의 판단과 권리 행사를 어떻게 지원할 수 있을지 논의한다.

2. 법률상 민감정보와 맥락에 따른 위험

「개인정보 보호법」 제23조 제1항은 사상·신념, 노동조합·정당의 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보 및 그 밖에 정보주체의 사생활을 현저히 침해할 우려가 있는 개인정보로서 대통령령으로 정하는 정보를 민감정보로 규정한다. 개인정보처리자는 원칙적으로 민감정보를 처리하여서는 아니 된다. 다만 제15조 제2항 각 호 또는 제17조 제2항 각 호의 사항을 정보주체에게 알리고 다른 개인정보의 처리에 대한 동의와 별도로 동의를 받은 경우, 또는 법령에서 민감정보의 처리를 요구하거나 허용하는 경우에는 처리할 수 있다.^[2]

학교명·위치·교우관계가 그 자체로 언제나 법률상 민감정보에 해당하는 것은 아니다. 다만 이러한 정보로 특정 개인을 알아볼 수 있거나 다른 정보와 쉽게 결합하여 개인을 식별할 수 있다면 개인정보에 해당하며, 목적에 필요한 최소한의 수집과 안전한 관리 등 관련 원칙의 적용 대상이 된다.^[3] 학교명에 귀가 시간과 자주 머무는 장소가 결합되면 특정 아동의 생활 동선이나 취약한 상황이 드러날 수 있다. 법정 민감정보인지에 대한 분류만으로 보호의 필요성을 모두 판단하기 어려운 이유이다. 실제 위험은 정보가 결합되는 방식과 이용 목적에 따라서도 달라질 수 있다.

같은 법 제22조의2 제1항은 만 14세 미만 아동의 개인정보를 처리하기 위하여 이 법에 따른 동의를 받아야 할 때에는 법정대리인의 동의를 받고 그 동의 여부를 확인하도록 한다. 이 요건은 모든 개인정보 처리에 일률적으로 적용되는 것이 아니라, 해당 처리에 법에 따른 동의가 필요한 경우에 적용된다. 같은 조 제3항은 만 14세 미만 아동에게 개인정보 처리와 관련한 사항의 고지 등을 할 때 이해하기 쉬운 양식과 명확하고 알기 쉬운 언어를 사용하도록 규정한다.^[4] 따라서 법정대리인에게 동의를 받는 절차와 아동에게 이해 가능한 방식으로 설명하는 일은 각각의 요건에 맞게 이루어져야 한다.

이러한 규정을 교육적 관점에서 해석하면, 보호자의 승인과 아동 자신의 이해를 함께 고려할 필요가 있다. 무엇이 수집되고 어떤 목적으로 사용되는지를 설명하는 과정에서 아동은 자신의 정보에 관한 판단에 참여할 수 있다. 설명을 제공했다는 사실만 확인할 것이 아니라, 아동이 궁금한 점을 묻거나 자신의 선택을 다시 살펴볼 수 있는 여건도 마련하는 것이 바람직하다. 이는 법조문에 명시된 요건을 넘어, 그 취지를 이용 경험에 반영하기 위한 실천적 제안이다.

만 14세라는 기준을 넘었다는 이유만으로 성인과 같은 이해와 통제 능력을 전제해서도 곤란하다. 고지의 적절성을 평가할 때에는 연령과 함께 설명의 난도, 이용 목적, 아동이 처한 상황을 고려할 필요가 있다. 청소년이 자신의 선택을 존중받는다라는 것은 모든 결과를 혼자 감당해야 한다는

뜻으로 해석되어서는 안 된다. 법률상 동의 요건을 충족하면서도, 이용자가 내용을 이해하고 필요한 도움을 얻을 수 있는 환경을 함께 마련해야 한다.

3. 자기노출에서 추론과 재이용으로 이어지는 위험

첫째, 대화 과정에서 이루어지는 자기노출은 정보 제공의 결과에 관한 충분한 숙고와 언제나 일치한다고 보기 어렵다. 아퀴스티 등의 연구는 개인정보 관련 판단과 행동을 설명하는 요소로 정보 제공의 결과와 개인 선호에 관한 불확실성, 맥락 의존성, 상업적·정부적 이해관계에 따른 영향 가능성을 제시한다.^[5] 이 연구는 아동의 챗봇 이용을 직접 검증한 실험은 아니지만 이를 대화 상황에 적용하면, 약관 동의 여부와 함께 실제 정보 제공이 이루어지는 조건을 살펴야 한다는 문제를 제기할 수 있다. 도움을 받으려는 과정에서 상세한 질문이 이어진다면, 이용자가 정보를 제공할 때 어떤 기대와 부담을 느낄 수 있는지도 검토할 필요가 있다.

AI 동반자가 정체성 탐색과 정서 표현을 위한 비교적 안전한 공간이 될 가능성을 논의하는 한편, AI 동반자 이용으로 또래·가족 등과 대면으로 교류하는 시간이 줄어들 가능성, 심리적 의존, 비현실적 관계 기대와 같은 위험을 검토한다. 또한 청소년의 기존 사회관계가 AI 동반자 이용에 영향을 줄 수 있다는 양방향적 관점을 제시한다.^[6] 이는 모든 챗봇 이용이 고립이나 개인정보 침해를 초래한다는 인과적 증거가 아니다. 친밀한 대화의 의미를 평가할 때에는 자기표현의 기회와 과도한 관여의 가능성을 함께 살피되, 서로 다른 이용자와 상황에 동일한 결과를 가정하지 않아야 한다.

둘째, 이용자가 직접 밝히지 않은 개인 속성이 텍스트로부터 추론될 수 있다. 스타브 등의 연구는 실제 Reddit 프로필 자료를 이용하여 대규모 언어모델이 위치, 소득, 성별 등의 속성을 추론할 수 있음을 보였다. 연구에서 검토한 텍스트 식별화와 모델 정렬 등의 완화 조치도 이러한 추론을 충분히 막지 못했다.^[7] 이 결과를 아동의 건강정보나 가정환경을 같은 정확도로 알아낼 수 있다는 증거로 일반화해서는 안 되지만 명시적인 식별정보를 삭제해도 표현 방식과 문맥에 개인에 관한 단서가 남을 수 있다는 점은 고려해야 한다. 직접 제공한 정보와 시스템이 새롭게 만들어 낸 추정 정보를 구분하여 살펴야 하는 이유가 여기에 있다.

셋째, 정보가 저장되고 재이용되는 과정에서는 최초 입력 이후의 위험을 검토해야 한다. 칼리니 등의 연구는 GPT-2를 대상으로 한 학습 데이터 추출 공격에서, 모델에 질의하여 학습 자료에 포함된 수백 개의 텍스트 시퀀스를 원문 그대로 추출할 수 있음을 보였다.^[8] 학습 데이터 추출이라는 기술적 위험과 실제 서비스의 보관·학습 정책을 구분해야 위험의 성격과 대응 범위를 정확하게 판단할 수 있다.

넷째, 대화에 포함된 정보는 여러 사람에 관한 것일 수 있으며, 기록의 영향이 장기간 이어질 가능성도 있다. 친구와의 갈등을 설명하면서 상담 내용이나 메시지 화면을 올리는 경우를 생각해 볼 수 있다. 발화자의 경험을 담은 설명 안에 상대방의 사생활이 함께 포함되는 것이다. 이때 정보보호의 범위를 계정 이용자 한 사람에게만 한정하면 대화에 등장하는 다른 사람의 권리를 놓칠 수 있다. 아울러 일시적인 고민이 지속적인 성격이나 위험 성향으로 분류된다면, 성장 과정

에 있는 아동이 과거의 한 장면으로 규정되는 낙인과 오분류의 위험을 검토해야 한다.

UNICEF의 「Guidance on AI and children」은 아동의 데이터와 프라이버시 보호를 아동 중심 AI의 핵심 요건으로 제시한다. 이에 따라 아동에 관한 정보를 수집·추론·이용하거나 장기간 보유하려는 경우에는, 아동의 최선의 이익과 정보 처리의 필요성·비례성·보유기간을 엄격하게 검토할 필요가 있다.^[9] 이러한 권고는 정보의 수집 시점뿐 아니라 시간이 흐른 뒤의 영향까지 고려할 필요성을 시사한다. 당시의 고민을 설명하는 데 필요했던 기록이 이후에도 같은 의미와 필요성을 갖는지는 별도의 문제이다. 아동의 성장과 변화 가능성은 데이터의 지속적인 보유와 해석을 검토할 때 중요한 윤리적 고려사항이 되어야 한다.

4. 아동의 권리를 서비스 설계와 운영 기준으로 구체화

앞서 살펴본 위험에 대응하려면 정보보호 원칙을 서비스의 실제 작동 방식으로 구체화해야 한다. 영국 정보위원회의 연령 적합 설계 규범은 아동이 이용할 가능성이 있는 온라인 서비스의 설계에서 아동의 최선의 이익을 주요 고려사항으로 삼도록 한다. 또한 아동의 최선의 이익을 고려하여 다른 기본 설정을 정당화할 충분한 근거가 있는 경우를 제외하고, 높은 수준의 개인정보 보호를 기본 설정으로 채택하도록 제시한다.^[10] 이러한 원칙은 국내 서비스의 보호 수준을 검토할 때에도 참고할 수 있다. 다음의 기준은 법률상 요건을 고려하면서 아동의 권리를 이용 경험에 반영하기 위한 설계·운영상의 제안이다.

첫째, 대화의 목적에 비추어 질문의 필요성을 판단하고, 불필요한 자기노출을 유도하지 않도록 해야 한다. 답변을 위해 추가 정보가 필요하다면 그 이유를 짧고 분명하게 설명하고, 이용자가 상세한 사생활을 밝히지 않아도 대화를 이어갈 수 있는 방법을 마련할 필요가 있다. 예컨대 학업 고민에 관한 일반적인 조언을 제공하는 상황이라면, 실명이나 학교명 대신 대략적인 상황만으로도 응답할 수 있는지 먼저 검토하는 것이다.

질문에 답하지 않거나 대화를 중단하는 선택도 자연스럽게 제시되어야 한다. 사적인 경험을 더 많이 말할수록 더 깊은 신뢰를 인정받는 것처럼 느끼게 하는 표현은 신중하게 검토할 필요가 있다. 설계 점검에서는 질문 하나의 적절성뿐 아니라 여러 차례의 질문이 이어질 때 요구되는 정보의 전체 범위도 살펴야 한다. 개별 질문은 평범해 보여도 대화가 누적되면서 수집의 범위가 넓어질 수 있기 때문이다. 이러한 점검은 아동에게 더 주의하라고 요구하는 안내를 실제 대화 구조의 개선으로 연결하는 과정이어야 한다.

둘째, 추론을 통해 생성한 정보를 별도의 검토 대상으로 삼아야 한다. 정서 상태나 가정환경을 추정하여 계정과 연결한다면, 그 추정이 서비스 제공에 필요한지, 정확성을 어떻게 평가할지, 어떤 결정에 활용할지를 구체적으로 정할 필요가 있다. 특히 추정치를 확인된 사실처럼 취급하여 이용자를 분류하지 않도록 불확실성과 오분류의 영향을 검토해야 한다. 법적 검토에서도 추론 정보가 특정 개인에 관한 개인정보인지, 나아가 건강 등에 관한 민감정보에 해당하는지를 내용과 처리 맥락에

따라 판단해야 한다.^[11] 추정이라는 형식만으로 정보 처리의 책임이 사라진다고 볼 수는 없다.

운영상으로는 추정에 사용한 정보와 추정 결과를 구분하고, 잘못된 분류에 이의를 제기하거나 재검토를 요청할 수 있는 경로를 마련하는 것이 바람직하다. 과거 기록의 해석을 바꾸거나 불필요해진 분류를 해제할 수 있는지도 점검해야 한다. 아동의 표현을 특정 성향의 확정적인 징표로 다루는 데에는 신중함이 요구된다. 한 번의 발화를 해석한 결과가 이후의 서비스 이용을 계속 좌우하도록 두기보다, 그 판단의 유효성과 필요성을 다시 살필 수 있어야 한다.

셋째, 저장·재이용·삭제의 범위를 이용자가 구분하여 이해할 수 있도록 해야 한다. 대화의 학습 이용 여부, 저장 기간, 사람에 의한 검토 가능성, 외부 사업자와의 공유 조건을 서비스별 정책과 기술적 설명에 근거하여 확인할 필요가 있다. 아동과 보호자에게 제공하는 안내에는 이러한 처리 중 무엇이 실제로 이루어지는지, 어떤 선택이 가능한지가 드러나야 한다. 이용 화면의 짧은 안내와 상세한 설명을 연결하여 필요한 정보를 쉽게 찾도록 하는 방식도 고려할 수 있다.

특히 화면에서 대화를 삭제했을 때 서버의 저장본이나 백업, 이미 학습에 사용된 데이터에 어떤 영향이 있는지를 구분하여 설명할 필요가 있다. 삭제 요청의 처리 결과와 남아 있는 데이터의 범위가 다르다면, 그 차이를 이해할 수 있게 알려야 한다. 단순히 삭제 기능이 있다는 사실만으로 이용자의 통제 가능성을 평가하기는 어렵다. 이용자가 원하는 조치를 요청할 수 있는지, 처리 결과를 확인할 수 있는지, 문의가 해결되지 않았을 때 다시 도움을 받을 수 있는지까지 살펴야 한다.

넷째, 제3자 정보의 처리 근거와 권리 보호 절차를 확인해야 한다. 입력자의 동의만으로 대화에 포함된 모든 사람의 개인정보 처리가 정당화된다고 가정해서는 안 된다. 적법성은 실제 정보주체, 처리 목적과 근거, 고지·동의의 필요성, 최소수집 원칙, 제공·공유 방식 등 구체적 사정에 따라 검토해야 한다. 이는 제3자 정보의 처리가 언제나 동의를 필요로 한다는 뜻이 아니라, 해당 정보에 적용되는 적법한 처리 근거를 확인해야 한다는 의미이다.^[12]

서비스 운영자는 이러한 기준의 이행을 확인할 책임과 절차도 정할 필요가 있다. 기능을 변경하거나 새로운 정보 활용 목적을 도입할 때에는 기존 안내와 보호 조치가 여전히 적절한지 재검토하고, 발견된 문제를 개선 과정에 반영하는 것이 바람직하다. 아동이 제기한 불편이나 이해하기 어려웠던 설명도 점검의 자료로 활용할 수 있다. 권리 보호가 선언에 머물지 않으려면 이용자의 경험을 듣고 운영 방식을 수정할 수 있는 절차가 지속적으로 작동해야 한다.

5. 자기표현과 미래의 자율성을 지키는 공동의 책임

아동·청소년의 AI 챗봇 이용을 둘러싼 정보보호는 사업자의 설계와 운영, 교육기관의 도입 판단, 가정의 일상적인 지원이 서로 연결될 때 실효성을 갖출 수 있다. 각 주체는 자신이 결정할 수 있는 범위에서 책임을 구체화해야 한다. 서비스의 내부 처리 방식을 알기 어려운 아동에게 모든 위험을 판별하도록 요구하거나, 보호자와 교사에게 그 부담을 그대로 이전하는 방식은 적절하지 않다. 기술을 제공하고 도입하는 주체의 설명과 아동의 선택을 지원하는 교육이 함께 이루어져야 한다.

교육기관은 챗봇의 교육적 활용 목적과 학생에게 요구되는 정보의 범위를 함께 검토할 필요가 있다. 특히 수업 참여나 과제 수행을 위해 가족관계, 건강 상태, 정서적 어려움과 같은 사적인 경험을 공개하도록 요구하지 않는 활동을 설계하는 것이 바람직하다. 가상의 사례를 활용하거나 개인을 식별하기 어려운 수준에서 상황을 설명하는 방법도 고려할 수 있다. 학생이 특정 서비스 이용이나 정보 제공을 부담스러워하는 경우에는 학습 목표에 도달할 수 있는 대체 수단을 마련하는 방안을 검토해야 한다.

개인정보 보호교육에는 자신의 정보와 타인의 정보를 구분하는 판단, 질문에 응답할 범위를 정하는 선택, 문제가 발생했을 때 도움을 요청하는 실천이 포함되어야 한다. 사례를 통해 무엇이 불편했는지, 어떤 정보를 덜 제공할 수 있었는지, 이미 입력한 내용에 대해 어떤 조치를 확인해야 하는지를 토론할 수 있다. 이러한 교육은 정답을 외우는 활동에 머물기보다, 자신과 타인의 사생활을 존중하는 이유를 설명하고 구체적인 상황에서 판단의 근거를 세우는 과정으로 구성할 필요가 있다.

가정에서는 아동의 발달 수준과 사생활을 존중하면서 정보 제공의 의미를 함께 살펴볼 필요가 있다. 원치 않는 정보를 입력했거나 낯선 질문에 불편함을 느꼈을 때, 비난을 걱정하지 않고 도움을 요청할 수 있는 관계를 마련하는 것이 중요하다. 보호자의 개입은 아동이 어떤 선택을 했는지 이해하고 필요한 조치를 함께 확인하는 방향으로 이루어지는 것이 바람직하다. 잘못된 판단의 결과를 혼자 감당하게 하기보다, 경험을 돌아보고 이후의 선택을 조정할 수 있도록 지원해야 한다.

보호 기준을 평가할 때에는 확인된 사실과 제한된 조건에서 관찰된 결과, 아직 검증이 필요한 가능성을 구분해야 한다. 새로운 근거에 따라 조치를 재검토하면서도, 위험의 불확실성을 이유로 설명과 책임을 미루어서는 안 된다. 궁극적으로 지켜야 할 것은 아동이 필요한 도움을 구하고 자신을 표현하면서도, 그 기록에 의해 미래의 선택을 부당하게 제약받지 않을 가능성이다. 오늘의 고민을 존중하고 내일의 변화를 허용하는 정보 환경을 마련하는 일이 우리 사회가 함께 감당해야 할 책무이다.

참고문헌

- [1] Nissenbaum, H., Privacy in context: Technology, policy, and the integrity of social life, Stanford University Press, 2010; Nissenbaum, H., Contextual integrity up and down the data food chain, Theoretical Inquiries in Law, 20(1), 2019, pp. 221-256.
- [2] 대한민국, 「개인정보 보호법」, 법률 제21445호(2026. 3. 10. 일부개정, 2026. 9. 11. 시행), 제23조 제1항.국가법령정보센터.
<https://www.law.go.kr/법령/개인정보보호법> (검색일: 2026. 09.15).
- [3] 대한민국, 「개인정보 보호법」, 제2조 제1호, 제16조 제1항, 제29조, 2026.
- [4] 대한민국, 「개인정보 보호법」, 제22조의2 제1항·제3항, 2026.
- [5] Acquisti, A., Brandimarte, L., and Loewenstein, G., Privacy and human behavior in the age of information, Science, 347(6221), 2015, pp. 509-514.
- [6] Sun, X., Wang, Y., and McDaniel, B. T., AI companions and adolescent social relationships: Benefits, risks, and bidirectional influences, Child Development Perspectives, 20(2), 2026, pp. 91-98.
- [7] Staab, R., Vero, M., Balunović, M., and Vechev, M., Beyond memorization: Violating privacy via inference with large language models, The Twelfth International Conference on Learning Representations, 2024.
- [8] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C., Extracting training data from large language models, 30th USENIX Security Symposium (USENIX Security 21), USENIX Association, 2021, pp. 2633-2650.
- [9] UNICEF Innocenti – Global Office of Research and Foresight, Guidance on AI and children, Version 3.0, UNICEF, 2025.12., pp. 22-23.
<https://www.unicef.org/innocenti/reports/policy-guidance-ai-children> (검색일: 2026. 09.15).
- [10] Information Commissioner's Office, Age appropriate design: A code of practice for online services, 2020.
<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/> (검색일: 2026. 09.24).
- [11] 대한민국, 「개인정보 보호법」, 제2조·제23조, 2026.
- [12] 대한민국, 「개인정보 보호법」, 제2조·제15조·제16조·제17조, 2026.

Privacy Report

2026 개인정보 이슈 심층 분석 보고서

『2026 개인정보 이슈 심층 분석 보고서』는
개인정보보호위원회 출연금으로 수행한
사업의 결과물입니다.

한국인터넷진흥원의 승인 없이 본 보고서의 무단전재나
복제를 금하며, 인용 출처 『2026 개인정보 이슈 심층 분석
보고서』를 밝혀주시기 바랍니다.

본 보고서의 내용은
한국인터넷진흥원의 공식 견해가 아님을 알려드립니다.

발행

발행일	2026년 10월
발행처	한국인터넷진흥원 개인정보예방팀 전남광주통합특별시 나주시 진흥길 9

Privacy Report

2026 개인정보 이슈
심층 분석 보고서

2026 Vol. 3