딥시크 활용과 개인정보보호: 위험 분석과 보안 대응 전략

1. **딥시크란?**

   딥시크(DeepSeek)는 2023년 중국 항저우에서 설립된 AI 스타트업이 개발한 대규모 언어모델(LLM) 및 생성형 AI 서비스로 코드 작성, 번역, 요약, 문서 작성 등 광범위한 과업을 수행한다. R1, V3 등 모델을 연속 공개하며 고성능을 내세웠고 특히 V3는 초대형 파라미터 규모를 강조하며 범용 작업 능력을 주장했다. 동시에 중국 정부의 검열 및 접근 가능성에 대한 우려가 제기되어 신뢰성 및 개방성 논쟁이 이어지고 있다.

2. **딥시크의 역할과 활용**

   - 범용 AI 작업: 자연어 처리 중심의 질의응답, 요약, 번역, 문서 초안 작성, 코딩 보조 등 다양한 작업에 투입되어 인간과 기계 간 협업 도구로 쓰인다.
   - 시장 파급력: 제한된 자원에서도 최적화로 높은 성능을 달성했다는 메시지로 글로벌 AI 경쟁 구도를 흔들며 오픈소스 친화 이미지를 곁들여 “AI 기술의 민주화”를 강조한다.

   이와 같은 역할은 비용 효율과 생산성 향상에 기여하지만 그 전제에 “무엇을 학습하고 어떤 데이터를 처리하는가”가 따라붙는다.

3. **개인정보보호 문제점**

   - 과도한 행동 데이터 수집 우려: 일반 AI도 IP 및 기기정보 등을 수집하지만 딥시크는 “사용자 키 입력 패턴 및 리듬(Keystroke Dynamics)” 등으로 개인을 식별 및 프로파일링이 가능하다는 보도가 나왔다. 이는 마케팅 및 추적에 악용될 소지가 크다는 지적이다.
   - 중국 내 저장과 정부 접근 가능성: 수집 데이터가 중국 서버에 저장되고 중국 법령상 공안 및 국가안전 등 사유로 정부 요구 시 기업이 협조해야 한다는 점이 국외 이전과 정부 접근 리스크를 키운다.
   - 국내 공공 및 대기업의 선제 차단: 한국의 주요 공공기관과 기업이 접근을 제한하고 개인정보보호위원회(PIPC)가 본사에 질의서 발송 및 기술분석, 해외 감독기구와 공조에 착수했다.

   이 문제군은 “민감정보·식별자·행동 데이터”가 조합될 때 개인의 프라이버시 침해와 2차 피해(스피어피싱 정밀화, 평판/차별 위험)로 비화할 수 있다.

4. **보안 위협 지형과 시나리오**

   - 데이터 유출: 학습 및 추론 과정에서 투입된 데이터(개인 혹은 기업 기밀)가 서비스

오류, 해킹, 설정 부주의로 외부에 노출될 위험이 있다. 입력 프롬프트나 업로드 파일이 재노출되는 2차 혹은 3차 피해가 포함된다.

- 모델 및 플랫폼 취약점 악용: 모델 탈취(Prompt Leakage), 파이프라인 공격, 공급망(서드파티 SDK 및 플러그인) 취약점으로 인한 무결성 훼손 가능성이 있다.
- 악용 증폭: 모델을 통한 악성코드, 랜섬웨어, 피싱 키트 자동생성, 범죄 고도화 등 AI를 이용한 사이버 범죄의 효율화 리스크가 존재한다.
- 국가안보 및 규모 리스크: 국가 배후 조직이 생성형 AI를 정보전 및 사이버공격에 활용하여 대규모 사회적 혹은 경제적 피해 가능성이 있다.

국제적으로는 NIST AI RMF, EU AI Act 등이 위험 정의 및 등급화와 관리 체계를 제시하며 교차 위험(Cross-cutting risk)까지 포괄하는 거버넌스 필요성을 강조한다.

5. **규제 및 컴플라이언스 환경**

- 국내 대응: 개인정보보호위원회가 본사 질의, 기술 분석, 국제 공조(영 ICO, 프 CNIL, 아일랜드 DPC) 등을 진행 중이다. 공식 외교 채널 및 한중 협력센터를 통한 소통도 병행하고 생성형 AI 안전 이용 가이드 배포를 예고했다.
- 국제 프레임워크:

1. NIST AI RMF 1.0: 위험을 “발생 가능성과 영향”의 결합으로 정의한다. 조직 차원의 신뢰 가능한 AI 개발 및 운영을 위한 프로세스를 제시한다.
2. EU AI Act: 위험 4단계(금지, 고위험, 특정 투명성, 최소 위험) 구분과 단계별 의무 부과로 안정성, 투명성, 책임성 확보를 지향한다.

이 환경은 조직이 타사 생성형 AI를 사용할 때 DPIA(영향 평가), 국외이전 적법성, 공급망 보증, 데이터 최소화 등 사전 통제의 중요성을 높인다.

6. **대응 방안**

1. **개인 사용자**

- 민감정보 미입력: 이름, 연락처, 계정 및 비밀번호, 주민번호, 여권, 운전면허번호, 금융정보, 위치 및 건강, 정치성향 등은 입력하지 않는다. 불가피하면 익명화 또는 마스킹 후 최소 항목만 사용한다.
- 약관 및 정책 확인: 데이터 수집 항목, 저장위치(국내 혹은 국외), 보존기간, 2차 활용 및 공유 여부, 모델 학습 재사용 여부를 필수로 확인한다.
- 대체 수단 활용: 데이터 국외이전 최소화 및 국내 규제 준수 명시 서비스, 온디바이스 및 브라우저 내 추론 등 데이터 잔존성이 낮은 도구를 우선 사용한다.
- 브라우저 및 네트워크 보호: 추적 차단, 스크립트 및 키로깅 방지 확장, 의심 도메인 차단, 2FA와 비밀번호 관리로 계정 탈취를 예방한다.
- 디지털 각성: AI 출력에 포함된 재식별 가능 정보(이름 및 이메일 등) 감지 시 공유

중단 및 삭제를 요청한다.

국내 보도 및 기관 권고는 민감정보 입력 회피와 서비스의 개인정보 처리 실태 확인을 반복적으로 강조한다.

2. **조직(기업 또는 기관)**

- 사전 통제(거버넌스):

a. 정책: 생성형 AI 사용 가이드, 금지 데이터 분류(비밀, 고유식별, 고객 PII), 프롬프트 및 파일 업로드 규칙을 수립한다.

b. 평가: 공급업체 실사(Vendor Due Diligence), DPIA, 국외이전 법적 근거, 계약서에 데이터 소유권, 학습 금지, 삭제 SLA, 침해 통지 의무를 명문화한다.

c. 규제 정합성: NIST AI RMF 기반 위험관리, EU AI Act의 고위험 요건(데이터, 로그, 투명성, 인간감독 등) 준거 설계한다.

- 기술 통제:

a. 접근 및 경계: AI 게이트웨이 및 프록시로 도메인 allowlist, 프롬프트 및 첨부파일 DLP 또는 PII 검출 및 차단, 토큰 혹은 API 키 보관을 금지한다.

b. 데이터 보호: 익명화 또는 가명처리, 프롬프트 레드락팅, 저장 금지 모드(Stateless) 우선, 암호화와 키 관리를 우선한다.

c. 대안 아키텍처: 온프레미스 및 가상사설망(VPC) 배포 모델, 데이터 레지던시 보장형, 파이어월드 가드레일 LLM을 사용한다.

d. 모니터링: 사용 로그 또는 감사 추적, 이상 탐지, 프롬프트 주입 및 데이터 추출 시그니처 룰을 확인한다.

- 운영 안전:

a. 레드팀 및 페네트레이션: 프롬프트 인젝션, 데이터 유출, 모델 탈취, 플러그인 및 SDK 공급망 공격 시나리오 정기 점검한다.

b. 교육 및 캠페인: 민감정보 미입력, 내부 데이터 외부 반출 금지, 승인된 채널 사용을 우선한다.

c. 사고대응: AI 관련 침해 유형 플레이북(격리, 토근 폐기, 벤더 통지 및 규제 신고), 삭제, 정정, 열람권 대응절차를 대비한다.

- 도입 기준:

a. 데이터 정책 투명성: 수집, 저장, 보존, 국외이전, 학습 재사용 불가 옵션을 명시한다.

b. 감사 가능성: 로그 접근 및 삭제 증빙, 독립 감사 보고서(SOC 2/ISO 27701 등)와 모델 보안 테스트 결과를 확인한다.

c. 정부 및 규제 리스크: 데이터 저장 위치, 현지 법률의 정부 접근 조항과 충돌 가능성을 평가한다.

국내 당국은 질의 및 기술분석, 국제 공조를 병행 중이며 조직은 그 결과와 가이드라

인을 정책, 프록시, DLP 등 기술 통제 업데이트에 즉시 반영하는 체계를 갖추어야 한다.