

데이터셋의 공정이용과 보상청구권

2023.8.25(금)

by 김윤명/법학박사/한림국제대학원대학교 교수



주요 내용

논의의 필요

데이터의 공정이용

학습데이터에 대한 보상청구권화

보상 모델 및 저작권법의 개정방안

결론

1. 논의의 필요(1)

- 구글, 메타 등 글로벌 빅테크 기업은 LLM을 구축하거나 생성형 AI형태로 자사 서비스에 이용하고 있으며, 국내에서는 LG 엑사온이 공개된 바 있음
- LLM을 구축하는 과정에서 수많은 데이터가 학습데이터로 이용되고 있으며, 그 과정에서 권리자의 허락을 받는 경우도 있지만 그렇지 않는 경우도 있음. 해외에서는 생성형 AI모델을 구축한 기업에 대해 저작권 침해 등을 이유로 집단소송(class action)이 제기되고 있음

1. 논의의 필요 (2)

- 사업자들은 데이터를 공개할 의무가 없기 때문에 사실상 구축된 데이터를 확인할 수 없음. 그렇지만, 일부 AI생성물을 보면 저작물과 유사한 결과물이 생성되고 있다는 점에서 데이터 구축과정에서 위법행위가 있음을 추론할 수 있음
- 인터넷에 공개된 수많은 게시물을 학습 데이터로 활용한다는 점에서 게시물 수량만큼 게시자도 많을 것으로 보이나, 게시자 개별적인 게시량은 그렇게 많지는 않을 것으로 생각됨
- 따라서, ‘소량의 이용’에 해당할 수 있다는 점과 공정이용(fair use) 규정에 따라 데이터의 TDM이나 기계학습이 공정이용이 인정될 가능성이 높음

1. 논의의 필요(3)

- (참고) 호주의 정치철학자로 '모든 것의 미래' 저자인 팀 던럽 박사는 2020년 국회에서 열린 '경기도 데이터 배당' 토론회에서 “데이터는 우리 모두가 소유하는 자원이기 때문에, 데이터 추출로 이익을 보는 어떠한 회사라도 그 혜택에 대한 대가를 지불해야 한다”면서 “데이터 추출 수익금을 보다 더 공정하게 재분배하는 것이야말로 활기찬 경제와 번영하는 사회를 보장하는 중요한 방법이 될 것”이라고 주장

2. 데이터의 공정이용(1)

1. 저작권의 성질

- 배타적 권리로서 사전 이용허락 필요
- 저작권자의 이용허락을 받는 과정에서 적지 않은 거래비용(transaction cost) 발생
- 예외적으로, 저작권자를 확인하기 어렵거나 방송과 같은 공익목적의 경우에는 보상청구권으로서 사후보상으로 이용 가능

2. 데이터의 공정이용 (2)

2. 저작권 행사의 제한

- 3단계 테스트 : “저작물의 통상적인 이용과 충돌하지 않고, 저작자의 합법적인 이익을 불합리하게 침해하지 않는 특정한 경우에 한하여 그러한 저작물의 복제를 허용하는 것”
- 저작권재산권의 제한
 - 저작권자의 허락없이 일정한 요건하에서 저작물을 이용할 수 있도록 함
 - 이용자의 권리가 아닌 일종의 ‘반사적 이익’으로 이해될 수 있음
- 공정이용
 1. 이용의 목적 및 성격
 2. 저작물의 종류 및 용도
 3. 이용된 부분이 저작물 전체에서 차지하는 비중과 그 중요성
 4. 저작물의 이용이 그 저작물의 현재 시장 또는 가치나 잠재적인 시장 또는 가치에 미치는 영향

2. 데이터의 공정이용 (3)

3. 데이터 공정이용과 보상

가. 데이터 공정이용

- TDM과 학습데이터의 구축
- 기계학습과 데이터

나. 보상청구권

- 학습데이터를 구축하는 과정에서 저작권 침해에 대한 법적안정성 확보 필요
- 보상청구권의 성격은 선사용 후 보상이라는 채권적 권리
- (유형) 저작재산권의 제한규정에 따른 청구권, 인접권자에 대한 청구권, 법정허락 제도 등

3. 학습데이터에 대한 보상청구권화 (1)

1. 왜?

- (문제제기) LLM사업자들의 데이터 공정이용과 과실의 독점은 공평한가?
- LLM사업자는 별도 비용없이 수익을 창출하게 됨으로써, 공정경쟁에 부합하지 않다는 주장이 가능
- (필요성) 인터넷상에 게시물을 올린 이용자에게도 보상이 돌아갈 수 있다는 점
- 반면, 사업자는 정당한 방식으로 데이터를 활용할 수 있다는 점에서 합법적 BM으로서 인정받을 수 있다는 점에서 법적 안정성을 확보할 수 있음

3. 학습데이터에 대한 보상청구권화 (2)

2. (해당 데이터) 모든 데이터를 대상으로 할 것인지?

- 어떻게 특정할 것인지?
- 신탁관리단체를 통한 확대된 집중관리 방안의 제시

3. 학습데이터에 대한 보상청구권화 (3)

3. (문제) 누구에게 보상할 것인가?

- 인터넷공간의 저작자 확인은 불가능한 영역
- 인터넷 게시물을 학습데이터로 구축하는 과정에서 권리 내지 권리자를 특정하기 어렵다는 점
- 또한, 게시자라고 하더라도 진정한 권리자인지를 확인하기가 쉽지 않음

3. 학습데이터에 대한 보상청구권화 (4)

4. (피청구권자) 학습데이터를 생성하는 기업으로 할 것인지?

- 학습데이터를 기반으로 AI모델을 구축하여 서비스하는 기업을 대상으로 할 것인지?
- 어떻게 사용한 것인지 여부를 특정할 수 있을지 고민
- 구글, 네이버 등 플랫폼 사업자에 대한 일정량의 청구권을 행사하는 방안
- 플랫폼을 갖추지 아니한 사업자에 대해서는 어떻게?

3. 학습데이터에 대한 보상청구권화 (5)

5. (보상금 회계처리) 기금이나 특별회계를 통한 세목 신설

- 데이터 활용하는 기업을 대상으로 발생하는 수익에 대한 법인세 성격
- 각각의 법령에 근거하여 설치 및 운용되고 있는 기금은 특정한 사유가 있을 때 법률로서 설치된다는 점에서 일반회계와 차이가 있고, 기금은 특별회계에 비해 운용상의 자율성이 상대적으로 크다고 할 수 있음
- 일반회계는 국가 고유의 일반적 재정활동에 사용되는 반면, 특별회계와 기금은 특정한 사업(목적) 또는 자금을 별도로 운영할 필요가 있거나, 특정한 세입(수입)과 세출(지출)을 직접 연계할 필요가 있을 때 설치함

3. 학습데이터에 대한 보상청구권화 (6)

6. 분배 및 기금 누적

- 문제는 소량의 데이터를 게시한 경우, 보상은 크지 않을뿐더러 게시자를 특정하여 분배하는 것은 불가능한 영역
- 분배액이 크지 않다는 것이, 자격이 없다거나 소유권이 없다는 주장의 근거가 되지 못함
- 분배의 실익이 크지 않기 때문에 누적적으로 관리하여, 공익목적으로 활용할 수 있는 방안 마련
- 기금, ‘특별회계’로 관리

3. 학습데이터에 대한 보상청구권화 (7)

7. (한계) 2중과세 여부

- 디지털세는 자원 보유국의 로열티처럼 다국적 플랫폼 기업이 특정 국가에서 획득한 지대에 대한 세금으로 이해. 자원 보유국은 자국의 천연자원 채굴에서 나오는 수익에 대하여 법인세와는 별도의 로열티, 즉 지대세를 채굴 기업에 부과.
- 아울러, 세계 각국의 표준 법인세 체계는 일상적 법인세에 더하여 초과 이윤과 같은 특수한 이익에 대한 부가적 과세 수단을 설정.
- 법인의 소득 중 비사업용 토지 등 양도소득이 있으면 그것에 대하여 법인세 기본세율에 더하여 10%의 추가 과세.
- 부동산 법인의 경우도, 사업소득에 대한 법인세와 더불어 부동산양도소득에 대한 세금을 추가로 납부해야 함. 즉, 법인세와 디지털세는 이중과세에 해당되지 않음.

4. 보상모델 및 저작권법 개정

1. 보상청구권 제도 검토

- 확대된 집중관리 형태?

2. 데이터보상금 청구권의 규정

- 저작권법의 개정 필요
- 공익 법인
- 기술적 실업 보상, 기본소득 등의 자원
- 특별회계로 누적적 적립

5. 결론

- 인터넷 등 다양한 정보를 활용하기 위해서는 저작권자 허락이 요구되나, 거래비용이 과다하기 때문에 개별적 이용허락은 실익이 크지 않음
- 공정이용이나 저작재산권 제한 규정에 근거하여, 데이터를 이용함으로써 사업자는 개별 게시자에게 보상은 하지 않으면서 과실을 독점하는 구조는 합리적으로 보기 어려움
- 플랫폼사업자의 거래비용을 줄이고, 권리자에 대한 보상체계로서 확대된 집중관리제도를 고려
- 개별적 배분은 쉽지 않아, 공익법인을 통한 기금 등으로 적립하여 배분 및 활용

고맙습니다.

질문은, digitallaw@naver.com 으로~