Plagiatul în timpul erei LLM

Student: Soare Robert Daniel
Facultatea: Automatică și Calculatoare - UPB
Master: Grafică, Multimedia și Realitate Virtuală (GMRV), Anul 2

Definiția plagiatului

Conform Dicționarului Explicativ Român, o definitie pentru acțiunea de **plagiat** este: „A-și însuși, a copia total sau parțial ideile, operele etc. cuiva, prezentându-le drept creații personale; a comite un furt literar, artistic sau științific” [1]. Prin legea nr. 206 din 27 mai 2004, plagiatul este definit ca fiind „preluarea integrală sau parțială, cu intenție sau fără intenție, a unei opere sau a unei părți dintr-o operă, fără respectarea drepturilor de autor sau a drepturilor conexe acestora, precum și preluarea integrală sau parțială, cu intenție sau fără intenție, a unei opere sau a unei părți dintr-o operă, cu respectarea drepturilor de autor sau a drepturilor conexe acestora, dar fără respectarea condițiilor impuse de legea care reglementează drepturile de autor și drepturile conexe acestora” [2].

Industria IT

Industria IT (*Information Technology*) este o industrie care se ocupă cu dezvoltarea, producția și comercializarea de produse și servicii bazate pe tehnologia informației. Aceasta este una dintre cele mai dinamice industrii din lume care într-o continuă creștere accelerată. În zilele noastre, aproape orice afacere are o prezență *online*, fie prin intermediul unui website propriu sau terț (platforme sociale precum *Facebook, Twitter, TikTok*).

Această creștere rapidă a condus și la dezvoltarea multor domenii precum *web development, mobile development, data science, machine learning, artificial intelligence* etc. Prin prisma acesti creșteri accelerate, a aparut și comunitatea *open source*. Comunitate care incurajează dezvoltarea de proiecte *open source* și contribuția la acestea. Aceste proiecte sunt disponibile public și oricine poate contribui la dezvoltarea lor.

Multe proiecte au fost construite în baza unor proiecte *open source* și au fost lansate ca produse comerciale. Un exemplu de astfel de proiect este *Android*, un sistem de operare pentru dispozitive mobile, care a fost construit pe baza proiectului *open source Linux*. Pe cât de frumoasă este mișcarea *open source*, aceasta a evidențiat și diferite probleme precum încălcarea licențelor, furtul de cod, copierea de cod, plagiatul etc [3].

De ce oamenii din industrie continuă să publice proiecte *open source* dacă există riscul ca acestea să fie copiate? Răspunsul este simplu: beneficiile depășesc pagubele. Aceste tipuri de proiecte sunt mai apreciate și bine venite decât cele care sunt cu sursă închisă. Astfel, *start-up*-urile și companiile mari pot beneficia de pe urma acestor proiecte, iar dezvoltatorii care contribuie la acestea pot fi recunoscuți de comunitate și pot fi contactați de companii pentru a lucra la proiecte comerciale [4].

Așadar, pentru această industrie, aceste acțiuni de copiere a codului sursă, de plagiat, de încălcare a licențelor, de furt intelectual sunt mai degrabă oportunități decât probleme [5]. Dar aceste acțiuni nu sunt binevenite în alte industrii precum industria medicală, industria militară, industria auto etc. În aceste industrii, aceste acțiuni sunt considerate infracțiuni și sunt pedepsite de lege.

Această lipsă a griji în privința acestor probleme și posibilitea colectării de date (voite sau nevoite – fapt care a cauzat și o reacție din partea Uniunii Europene: *GDPR* [6]) în scopuri de cercetare sau comerciale a dus la nașterea marii revoluții care a schimbat lumea: *Large Language Models*.

Era *Large Language Models*

Odată cu explozia de tehnologiilor bazate pe *Large Language Models* (LLM). Aceste tehnologii sunt bazate pe modele de învățare automată care au fost antrenate pe un set de date foarte mari. Aceste modele pot fi folosite pentru a genera texte, cod, imagini, sunete etc. Un exemplu de astfel de tehnologie este *GPT-4* [7]. O problemă care era de multe ori ignorată a intrat în centrul atenției: furtul de cod.

Aceste tehnologii aa facut ca furtul intelectual să fie foarte ușor și invizibil, utilizatorull neștiind că au copiat codul de undeva. Un exemplu de astfel de proiect este *Copilot* . Acesta este un produs dezvoltat de *GitHub* și *OpenAI* care poate genera cod pe baza unui text. Acesta a fost lansat în luna iulie a anului 2021 și a fost criticat de comunitatea *open source* pentru că încurajează furtul de cod [8].

Dar tehnologii precum GPT-4 nu au influențat doar industria IT, aceastea încep să schimbe complet lumea. Marile corporații precum *Google, Microsoft, Facebook* etc. au început să investească în aceste tehnologii. Acestea au început să fie folosite în industria medicală, în industria auto, în industria militară etc. Pe data de 10 Mai 2023, Google anunță integrarea tehnologiei PaLM 2 (*Pathways Language Model 2*) în mai toate serviciile sale – arătând că *LLM* este viitorul și în curând prezentul pentru toți. În ciuda criticilor și reacților negative din partea comunității IT, marile companii merg mai departe cu această tehnologie, fiecare încercând să fie primul care o integrează cel mai bine în produsele sale.

Fară nicio exagerare, intrăm într-o nouă eră în care sancțiunile pentru acțiuni precum furtul intelectual, plagiatul, încălcarea licențelor trebuiesc revizuite și actualizate. Această tehnologie este prea valoroasă pentru a fi interzisă, dar în același timp este prea periculoasă pentru a fi lăsată să fie folosită în mod abuziv.

Guvernul Statelor Unite ale Americii a convocat marii reprezentanți ale acestor tehnologii la o discuție de stabilire a unor noi strategii prin care SUA să poată beneficia de pe urma acestei tehnologii, dar în același timp să poată proteja drepturile de autor și drepturile conexe acestora [9]. De asemenea, Uniunea Europeană lucrează la un proiectul de lege The Artificial Intelligence Act care va reglementa utilizarea acestor tehnologii [10].

Concluzie

Large Language Models schimbă complet modul cum intereacționăm cu tehnologia. Faptul că aceasta înțelege limbajul natural și poate genera texte, cod, imagini, sunete este un lucru incredibil. Dar acestă unealtă incredibilă creează și probleme de etică în privința informațiilor pe care le furnizează.

Soluția la această problemă este adaptarea legilor și a sancțiunilor pentru a putea proteja drepturile de autor și drepturile conexe acestora, dar în același timp să se poată beneficia de pe urma acestei tehnologii.

Între timp, cursa pentru perfecționarea acestor tehnologii continuă, iar guvernele și companiile încearcă în mod activ să profite de pe urma lor. Tot în timp vom vedea cum creatorii de conținut vor fi recompensați pentru munca lor, muncă care este acum mai ușor de copiat.

Bibliografie

[1] “Plagiat.” https://dexonline.ro/definitie/plagiat/definitii

[2] “LEGE nr. 206 din 27 mai 2004.” https://legislatie.just.ro/Public/DetaliiDocument/52457

[3] Jaideep Reddy, “The consequences of violating open source licenses.” https://btlj.org/2015/11/consequences-violating-open-source-licenses/

[4] Lakhani, Karim R., and Robert G. Wolf, “Why hackers do what they do: Understanding motivation and effort in free/open source software projects.,” Perspectives on Free and Open Source Software, 2003.

[5] Yevgeniy (Jim) Brikman, “Why the best companies and developers give away almost everything they do.” https://www.ycombinator.com/library/56-why-the-best-companies-and-developers-give-away-almost-everything-they-do

[6] “General data protection regulation.” https://gdpr-info.eu/

[7] “Gpt-4.” https://openai.com/research/gpt-4

[8] “Github copilot litigation.” https://githubcopilotlitigation.com/

[9] “FACT SHEET biden-harris administration announces new actions to promote responsible AI innovation that protects americans’ rights and safety.” https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/

[10] “The artificial intelligence act.” https://artificialintelligenceact.eu/