

Šta Anthropicov novi "voden žig" zaista znači u kontekstu EU AI Act

Anthropic je nedavno potvrdio da Claude modeli lansirani počev od 2. avgusta 2026. godine u tekst koji generišu ugrađuju nevidljivi "voden žig". O ovoj novini se mnogo izveštavalo, ali ne uvek precizno.



UROŠ RAJIĆ

STARIJI SARADNIK



Pojedini mediji predstavili su je kao veliki iskorak u detekciji AI sadržaja. Drugi su je opisali kao rizik po privatnost. Nijedno od ova dva tumačenja nije sasvim tačno, a ta razlika je važna za svakoga ko razvija proizvod zasnovan na Claude-u ili ga koristi za kreiranje sadržaja koji dolazi do korisnika u EU, uključujući kompanije iz Srbije i šireg regiona Balkana.

Šta ova oznaka zapravo predstavlja

Prema sopstvenoj dokumentaciji Anthropic, oznaka je utkana u sam generisani tekst — nije dodata u vidu metapodataka niti skrivenih karaktera. Ona ne menja značenje, kvalitet niti čitljivost odgovora. Budući da je sastavni deo samog teksta, može opstati nakon kopiranja i lepljenja, a u određenoj meri i nakon naknadnog uređivanja. Primenjuje se na nivou modela, pa se pojavljuje bez obzira na to preko kog Claude proizvoda ili interfejsa je tekst generisan — bilo da je reč o API-ju, Claude.ai, Claude Code-u ili proizvodu treće strane izgrađenom na tom modelu. Anthropic ovu funkcionalnost uvodi globalno, a ne samo za sadržaj namenjen korisnicima u EU. Zbog toga će i kompanije u Srbiji i na Balkanu koje Claude koriste za tržišta van EU dobijati ovako označene izlazne rezultate, iako obaveza koja je podstakla uvođenje ove mere potiče iz prava EU.

Anthropic ovo dopunjuje još jednom tehnikom, namenjenom datotekama. Kada Claude generiše podržani tip datoteke, kao što su SVG, PNG ili JPG, dodaje potpisane metapodatke o poreklu sadržaja u skladu sa standardom C2P — istim otvorenim standardom koji se širom industrije koristi za beleženje porekla određenog sadržaja.

Oba mehanizma uvedena su radi ispunjavanja jedne konkretne pravne obaveze. Anthropic je pristupio Kodeksu prakse EU Acta o veštačkoj inteligenciji koji se odnosi na transparentnost sadržaja generisanog veštačkom inteligencijom, a ovo je način na koji Anthropic tu obavezu pretvara u funkcionalno tehničko rešenje.

Kako se ovo uklapa u EU AI Act

Član 50. [EU AI Acta](#) lako je potceniti, jer ne predviđa zahtevne postupke ocenjivanja usaglašenosti koji se primenjuju na visokorizične sisteme. Međutim, njegov domet širi je nego kod gotovo bilo koje druge odredbe Acta, zato što se primenjuje na svaku organizaciju koja se nađe u jednoj od četiri situacije, bez obzira na to da li njena upotreba veštačke inteligencije uopšte spada u kategoriju visokog rizika.

Vredi te četiri situacije jasno navesti, jer se obaveze odnose na različite aktere u lancu:

- Pružalac AI sistema koji neposredno komunicira sa ljudima, kao što je četbot, mora sistem projektovati tako da korisnici znaju da komuniciraju sa veštačkom inteligencijom.
- Pružalac sistema koji generiše sintetički audio, slikovni, video ili tekstualni sadržaj — uključujući modele opšte namene kao što je Claude — mora taj izlazni sadržaj označiti u mašinski čitljivom formatu koji omogućava da se utvrdi da je generisan veštačkom inteligencijom. Upravo na ovu obavezu odgovara Anthropicova oznaka.
- Korisnik sistema, odnosno deployer, koji koristi AI za prepoznavanje emocija ili biometrijsku kategorizaciju mora o tome obavestiti osobe koje su takvoj obradi izložene.
- Deployer koji objavljuje deepfake sadržaj ili tekst generisan veštačkom inteligencijom sa namerom da informiše javnost o pitanju od javnog interesa mora jasno naznačiti da je sadržaj generisan veštačkom inteligencijom, osim ako ga je čovek pregledao i preuzeo uredničku odgovornost za njega.

Kodeks prakse kojem je Anthropic pristupio predstavlja dobrovoljni instrument osmišljen da ove obaveze operacionalizuje, naročito obavezu označavanja iz druge navedene situacije. Evropska komisija i Evropski odbor za veštačku inteligenciju potom su ocenili da je Kodeks adekvatan za dokazivanje usklađenosti.

Komisija, međutim, precizno naglašava da samo pridržavanje odredaba Kodeksa ne predstavlja konačan dokaz usklađenosti. Ipak, on potpisnicima pruža značajan stepen pravne izvesnosti u pogledu načina na koji će regulatorna tela ocenjivati njihove prakse.

Za one koji prate rokove, važno je napomenuti da je paketom AI Omnibus promenjen samo jedan deo člana 50. Obaveza mašinski čitljivog označavanja iz člana 50(2) odložena je sa 2. avgusta 2026. na 2. decembar 2026. godine, i to samo za generativne AI sisteme koji su već bili stavljeni na tržište pre 2. avgusta 2026. Obaveze deployera u pogledu obaveštavanja i transparentnosti iz člana 50(1), (3) i (4) nisu odložene i počele su da se primenjuju 2. avgusta 2026, kako je prvobitno i bilo predviđeno. Anthropic je, uprkos tome, odlučio da označavanje ugradi u svoje modele prema prvobitnom vremenskom planu.

Šta ova oznaka ne može da uradi

Upravo se ovaj deo najčešće izgubi u naslovima, a najvažniji je za svakoga ko namerava da se na ovu tehnologiju osloni kao na instrument usklađenosti. Anthropic u sopstvenoj dokumentaciji otvoreno govori o ograničenjima sistema i ta ograničenja vredni ponoviti bez ublažavanja.

Detektovana oznaka jeste signal, ali nije dokaz. On ukazuje na to da je Claude možda obrađivao određeni sadržaj. Ne dokazuje da ga je Claude od početka napisao. Na primer, neko je Claude mogao koristiti samo za lekturu, prevodjenje ili sažimanje dokumenta koji je inače napisao čovek, a izlazni tekst bi i dalje mogao da sadrži oznaku. Sadržaj se nakon Claude-ove obrade takođe može dodatno uređivati ili kombinovati sa drugim sadržajem, što menja ono za šta se zapravo može reći da "vođeni žig" potvrđuje.

I obrnuto je podjednako tačno. Odsustvo detektovanog žiga ne znači da je sadržaj napisao čovek. Opsežno prepravljanje može ukloniti signal.

Kratki odlomci možda ne sadrže dovoljno teksta za pouzdanu detekciju. Snimci ekrana i konverzije između različitih formata datoteka često u potpunosti uklanjaju metapodatke. Pored toga, modeli objavljeni pre uvođenja ovog sistema označavanja uopšte ne sadrže takvu oznaku, budući da Anthropic još radi na naknadnom uvođenju ove funkcionalnosti u starije modele i nije se obavezao na konkretan vremenski plan za pojedinačne proizvode. Regulatorni krajnji rok je, međutim, 2. decembar 2026. godine — rok iz Omnibus paketa do kojeg pružaoci postojećih generativnih AI sistema moraju uvesti mašinski čitljivo označavanje.

Ništa od ovoga ne znači da je ova oznaka beskorisna. Naprotiv, on predstavlja stvaran i koristan signal za platforme i istraživače koji razvijaju alate za detekciju. Ali on je deo infrastrukture za usklađenost sa propisima, a ne dokaz autorstva, i kao takav ne bi trebalo da se tretira ni u ugovoru, ni u sporu, niti u podnesku regulatornom organu.

Zašto je ovo važno ako Srbija nije članica EU

Domet AI Acta nije ograničen na kompanije sa sedištem u Evropskoj uniji. On se primenjuje i na pružaocima i deployere van Unije kada se izlazni rezultat njihovog AI sistema stavlja na tržište EU ili koristi u EU. Kompanija sa sedištem u Beogradu koja vodi četbot zasnovan na AI-ju namenjen korisnicima u EU, generiše marketinški sadržaj usmeren ka potrošačima u EU ili proizvodi sadržaj informativnog karaktera koji se objavljuje za publiku u EU može potpasti direktno pod obaveze deployera iz člana 50, čak i ako uopšte nema poslovno prisustvo u EU.

Upravo ovde razlika između onoga što čini Anthropic i onoga što moraju učiniti njegovi korisnici dobija praktičan značaj. Anthropicov žig služi ispunjavanju sopstvene obaveze Anthropica kao pružaoca iz člana 50(2). On ne ispunjava vašu obavezu kao deployera prema članu 50(1), (3) ili (4), ukoliko vaša upotreba Claude-a potpada pod neku od tih kategorija.

To su zasebne obaveze transparentnosti, koje moraju biti vidljive i razumljive ljudima, a činjenica da je proizvod izgrađen na modelu koji u pozadini neprimetno označava svoj izlazni sadržaj ne izvršava nijednu od tih obaveza umesto vas.

Šta to znači u praksi

Za svaku kompaniju koja razvija proizvod na osnovu Claude API-ja ili ga koristi za generisanje sadržaja namenjenog korisnicima, tri stvari vredni proveriti sada, a ne tek kada to zatraži regulator. Prvo, utvrdite koja se kategorija iz člana 50. zaista primenjuje na vaš konkretan slučaj upotrebe, jer četbot, alat za generisanje sadržaja i biometrijska funkcionalnost aktiviraju različite obaveze transparentnosti i zahtevaju različita projektna rešenja.

Drugo, nemojte pretpostaviti da "vođeni žig" ugrađen u osnovni model ispunjava vašu obavezu ako vaš proizvod zahteva vidljivo obaveštenje o upotrebi veštačke inteligencije ili označavanje deepfake sadržaja. Te funkcionalnosti moraju biti ugrađene u vaš proizvod; ne mogu se jednostavno "naslediti" od modela.

Treće, vodite sopstvenu dokumentaciju o tome kako se sadržaj generisan veštačkom inteligencijom kreće kroz vašu organizaciju, jer ograničenja ove oznake znače da on sam po sebi neće moći pouzdano da odgovori na pitanja o autorstvu ako naknadno dođe do spora ili revizije. Pored ove tri provere, još jedan datum vredni obeležiti u kalendaru: 2. februar 2027. godine. To je rok predviđen Kodeksom prakse do kojeg pružaoci treba da uspostave interoperabilno rešenje za detekciju žigova. Time se ne uvodi nova obaveza za deployere. Ipak, to znači da bi prethodno opisani sistem — koji pruža „signal, a ne dokaz“ — trebalo da postane pouzdaniji kada odgovarajući alati za detekciju budu dostupni, pa će tada vredeti ponovo razmotriti njegove mogućnosti i ograničenja.

Član 50. stavlja u povoljniji položaj organizacije koje transparentnost tretiraju kao zahtev koji se ugrađuje u dizajn proizvoda od samog početka, a ne kao nešto o čemu se razmišlja naknadno. Rokovi su sada dovoljno blizu da ovo pitanje zaslužuje temeljnu analizu, a ne nagađanje.