WRZ 2

# AI Governance & Regulatory Update

## NAJWAŻNIEJSZE WYDARZENIA

### Komisja zapowiada dwie grupy robocze dla sygnatariuszy Kodeksu przejrzystości treści AI

Policy Development EU · czytaj całość ↓

Komisja Europejska zapowiedziała uruchomienie we wrześniu 2026 r. dwóch grup roboczych dla sygnatariuszy Kodeksu Postępowania ws. przejrzystości treści generowanych przez AI. Grupy mają służyć wymianie dobrych praktyk, zbieraniu informacji zwrotnej o wdrażaniu środków przejrzystości oraz rozwijaniu skutecznego stosowania Kodeksu.

### Dostawcy AI ujawniają różne sposoby wdrażania wymogów przejrzystości i znakowania treści z Art. 50

Market Response EU · czytaj całość ↓

W drugiej połowie sierpnia coraz wyraźniej było widać, że główni dostawcy realizują wymogi przejrzystości z Art. 50 różnymi środkami technicznymi i organizacyjnymi. Najbardziej szczegółowy opis przedstawił Anthropic: przyszłe modele Claude mają stosować niewidoczny watermark tekstowy, a obsługiwane pliki otrzymują Content Credentials w standardzie C2PA; równolegle Google, Meta, OpenAI i Microsoft komunikowały własne działania związane z transparentnością i zgodnością.

## REGULATORY NEWS

### Komisja zapowiada dwie grupy robocze wspierające wdrażanie Kodeksu Postępowania ws. przejrzystości treści generowanych przez AI

Policy Development EU · 2026-08-20

**Co się wydarzyło:** Komisja Europejska potwierdziła, że sygnatariusze Kodeksu Postępowania ws. przejrzystości treści generowanych przez AI zostaną zaproszeni do współpracy w dwóch grupach roboczych, których uruchomienie zaplanowano na wrzesień 2026 r.

**Kontekst:** Kodeks opisuje środki wspierające dostawców i podmioty wdrażające generatywne systemy AI w realizacji obowiązków dotyczących znakowania i etykietowania treści generowanych przez AI. Według Komisji do końca lipca Kodeks podpisało około 190 organizacji.

Cele dwóch grup roboczych obejmują:

- wymianę dobrych praktyk branżowych,
- zbieranie informacji zwrotnej dotyczącej wdrażania środków przejrzystości,
- rozwijanie stanu techniki i skutecznego stosowania Kodeksu.

**Status:** Uruchomienie grup było zapowiedziane na wrzesień 2026 r. Na dzień cut-off briefu ich prace nie były jeszcze rozpoczęte w materiale źródłowym wykorzystanym do tego wydania.

**Źródła:** Komisja Europejska – Strong backing for the Code of Practice on Transparency of AI-generated Content

## MARKET & COMPLIANCE

### Dostawcy AI ujawniają różne techniczne i organizacyjne sposoby wdrażania wymogów przejrzystości z Art. 50

Market Response EU · 2026-08-14..2026-09-01

**Co się wydarzyło:** W sierpniu 2026 r. pojawiły się bardziej konkretne informacje o tym, jak poszczególni dostawcy zamierzają realizować obowiązki dotyczące identyfikowalności i przejrzystości treści generowanych przez AI. Zestawienie Axios z 28 sierpnia pokazało, że nie powstaje jeden jednolity model wdrożenia: Anthropic opisał watermark tekstowy dla Claude, Google i Meta komunikowały rozwój narzędzi przejrzystości i watermarkingu, OpenAI wskazywało na podsumowania danych treningowych oraz osadzane sygnały pochodzenia, a Microsoft akcentował zmiany w wewnętrznym governance i zarządzaniu ryzykiem.

Najbardziej szczegółowe informacje techniczne opublikował Anthropic. Dostawca zapowiedział, że przyszłe modele Claude będą generować tekst zawierający niewidoczny dla czytelnika watermark, wdrażany w związku z wymogami EU AI Act. Strona Anthropic została 1 września zaktualizowana o aktualny status API do wykrywania watermarku, udostępnianego w trybie private preview uprawnionym organizacjom.

**Kontekst:** Watermark tekstowy Anthropic nie polega na dodawaniu ukrytych znaków ani dodatkowych metadanych do tekstu. Mechanizm jest oparty na wersji podejścia SynthID-Text i modyfikuje źródło losowości wykorzystywane przy wyborze kolejnych słów, tak aby w sekwencji wygenerowanego tekstu powstał wzorzec możliwy do wykrycia przy użyciu odpowiedniego klucza. Dla czytelnika tekst oznaczony i nieoznaczony ma być nierozróżnialny.

Anthropic wskazuje jednocześnie ograniczenia rozwiązania:

- watermark pozwala ocenić prawdopodobieństwo udziału Claude w powstaniu lub przetworzeniu tekstu, ale nie dowodzi autorstwa;
- detekcja jest słabsza przy krótkich fragmentach, tekstach silnie faktograficznych, korekcie istniejącego tekstu i kodzie, gdzie model ma mniej swobody wyboru słów;
- lekkie redagowanie prawdopodobnie nie usuwa sygnału całkowicie, natomiast pełne przepisanie tekstu może go usunąć;
- watermark nie zawiera informacji pozwalających zidentyfikować użytkownika, organizację ani konkretną rozmowę.

Dla plików mechanizm jest inny. Gdy Claude generuje obsługiwany typ pliku, m.in. PNG, JPG lub SVG, Anthropic dodaje kryptograficznie podpisane Content Credentials w metadanych, wykorzystując otwarty standard C2PA. Dostawca wyraźnie odróżnia ten mechanizm od watermarku tekstowego: C2PA zapisuje informację o pochodzeniu pliku w metadanych, podczas gdy watermark tekstowy jest wzorcem wynikającym z samego procesu generowania tekstu.

Anthropic zapowiedział również rozszerzanie watermarkingu na modele wprowadzone przed 2 sierpnia 2026 r. i stosowanie mechanizmu globalnie przy jego uruchomieniu. API detekcyjne ma być początkowo dostępne m.in. dla uprawnionych regulatorów, organów ścigania, mediów, fact-checkerów, badaczy, organizacji edukacyjnych i części przedsiębiorstw objętych obowiązkami weryfikacyjnymi.

Szerszy obraz rynku pozostaje zróżnicowany. Według Axios:

- Google i Meta deklarowały rozwój oraz stosowanie narzędzi przejrzystości i watermarkingu;
- Meta wskazywała również na współpracę z EU AI Office nad praktycznym i interoperacyjnym podejściem do zgodności;
- OpenAI publikowało podsumowania danych treningowych oraz stosowało osadzane „provenance signals”;
- Microsoft opisywał swoje dostosowanie do AI Act przede wszystkim przez zmiany wewnętrznego governance i zarządzania ryzykiem.

Informacje te pokazują różne warstwy realizacji obowiązków – od sygnału osadzonego w generowanym tekście, przez Content Credentials dla plików i inne sygnały pochodzenia, po rozwiązania governance. Nie oznacza to, że wszystkie wskazane mechanizmy są równoważne technicznie albo że każdy z dostawców wdrożył już pełny zestaw środków.

**Status:** Anthropic publicznie opisał mechanizm watermarkingu tekstowego i C2PA dla obsługiwanych plików oraz udostępnienie detection API w private preview. Watermarking został zapowiedziany dla przyszłych modeli Claude, a dla starszych modeli pozostawał w trakcie wdrażania. W odniesieniu do Google, Meta, OpenAI i Microsoft w tym wydaniu odnotowano działania opisane przez Axios; brief nie przesądza na tej podstawie, że którykolwiek z tych dostawców osiągnął pełną zgodność z Art. 50 ani że stosuje konkretną technologię poza tym, co zostało wskazane w źródle.

**Źródła:** Anthropic – How Claude's text watermark works · Komisja Europejska – Strong backing for the Code of Practice on Transparency of AI-generated Content · Axios – Europe's AI Act gets real · TechCrunch – Anthropic says it will watermark text generated by its AI models

## UPCOMING

- **2026-09** – Planowane uruchomienie dwóch grup roboczych dla sygnatariuszy Kodeksu Postępowania ws. przejrzystości treści generowanych przez AI · Komisja Europejska

- **2026-12-02** – Dla systemów AI wprowadzonych na rynek przed 2 sierpnia 2026 r. kończy się ograniczony okres przejściowy dotyczący obowiązku znakowania i wykrywania treści generowanych przez AI z Art. 50 ust. 2 · Komisja Europejska
- **2026-12-02** – Zaczynają obowiązywać zakazy dotyczące generowania lub manipulowania materiałami intymnymi bez zgody oraz materiałami przedstawiającymi seksualne wykorzystywanie dzieci (CSAM) · Komisja Europejska
- **2027-12-02** – Zaczynają obowiązywać przepisy dotyczące systemów AI wysokiego ryzyka wymienionych w Załączniku III do AI Act · Komisja Europejska
- **2028-08-02** – Zaczynają obowiązywać przepisy dotyczące systemów AI wysokiego ryzyka wbudowanych w produkty regulowane · Komisja Europejska

---

AI Opracowano z użyciem AI. Treść zweryfikował i zatwierdził **Łukasz Lewicki**.
ai.risk.compliance@gmail.com linkedin.com/company/ai-risk-compliance