Rozprzestrzenianie fałszywych wiadomości lub zamierzona dezinformacja przy użyciu AI

… i jak się bronić.

Prof. dr hab. Grażyna Szpor
Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie

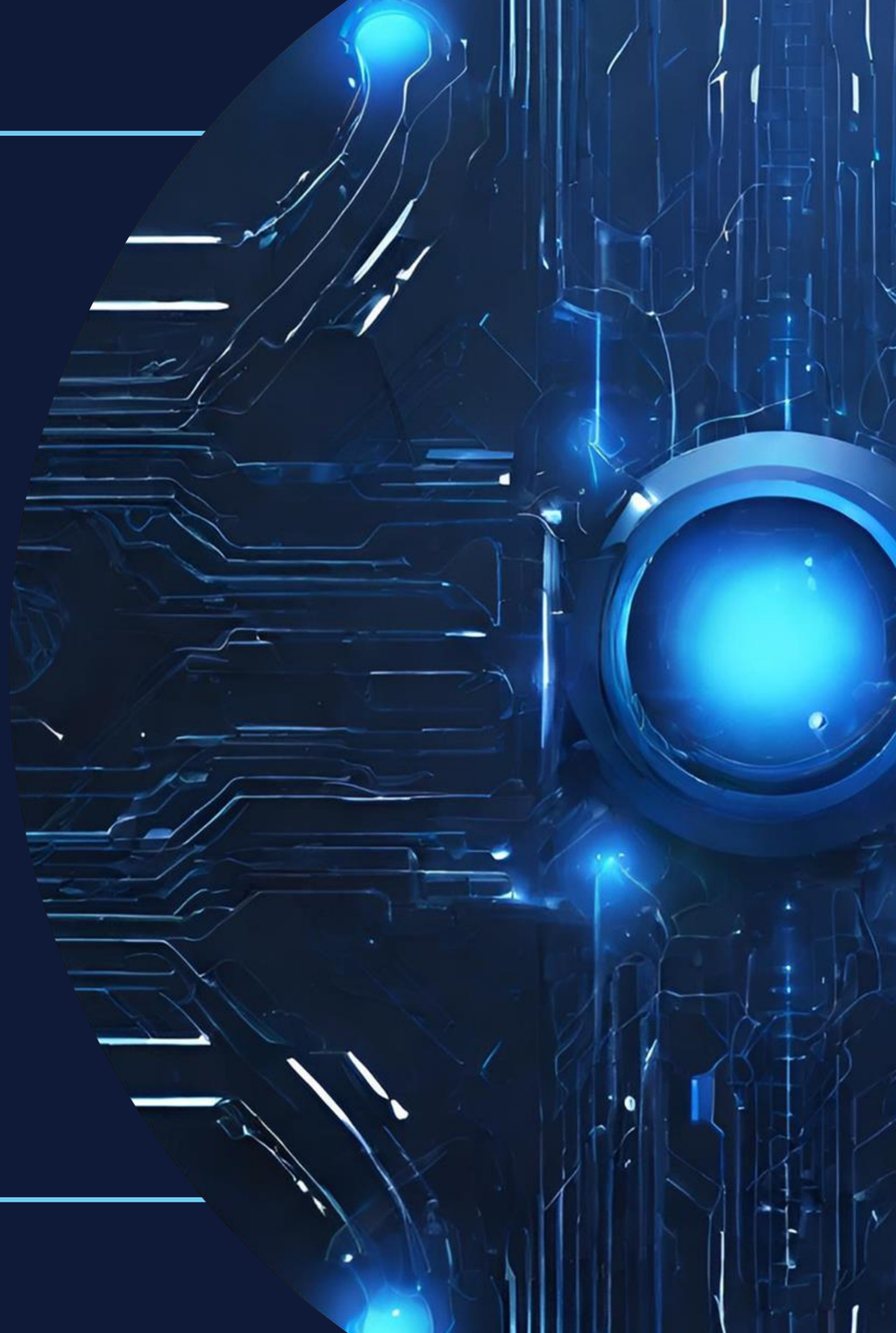
Agenda

- Czy dezinformacja to rodzaj informacji czy jej przeciwieństwo (odwrotność)?
- Jak sztuczna inteligencja zmienia dezinformację?
- Czy jesteśmy odporni na dezinformację?
- Czy instytucjonalna ochrona przed dezinformacją jest dostateczna?
- Jak wzmacniać samoobronę przed dezinformacją?

1

Dezinformacja

Rodzaj informacji,
czy przeciwieństwo informacji?



Wiadomość (Nachricht)

- Składa się z danych - znaków nadających się do przetwarzania na nośnikach fizycznych.
- Może być prawdziwa lub fałszywa (Fake news).
- W informatyce przyjmuje się, że wiadomość niesie informację, jeżeli redukuje entropię (tzn. zmniejsza chaos, niepewność); jeżeli tak nie jest, to dane mają zerową wartość informacyjną.
- Warto unikać nazywania fałszywej wiadomości (fake news) fałszywą informacją, bo trudniej wtedy zwalczać fake newsy.

Informacja

- Termin często używany ale niejednoznaczny, różnie definiowany w różnych naukach.
- W interdyscyplinarnym ujęciu informacja jest dobrem zmniejszającym niepewność, chaos (entropię,) przekazywanym przy użyciu różnych sygnałów.

  Wielka Encyklopedia Prawa (Tom XXII Prawo Informatyczne. Warszawa 2021, s. 204).
- Prawo człowieka do informacji to w istocie prawo do zmniejszania niepewności swojej i innych ludzi.

Dezinformacja

- Przedrostek de (dez)- (dis) wyrazów złożonych oznacza zaprzeczenie, pozbawienie lub odwrotność tego, co nazywa drugi człon złożenia”.

  Wielka Encyklopedia Prawa (Tom XXII Prawo Informatyczne. Warszawa 2021, s. 110).

- Dezinformacja jako odwrotność informacji jest złem (brakiem dobra, brakiem cechy **prawdziwości** przysługującej **wiadomości** odpowiednio do jej natury zwiększającym niepewność, chaos (entropię).

  A. Podsiad Słownik terminów i pojęć filozoficznych Warszawa 2001, s. 962-963.

- Dezinformacja to zamierzone wprowadzanie w błąd, podstęp.

Dezinformacja

- Prawie nigdy nie usiłuje zakryć prawdy.
- Daje przeciwnikowi argumenty, aby wierzył w to, w co chce wierzyć.
- Jest dla przeciwnika pretekstem wystarczającym do zignorowania dowodów przeszkadzających mu w wyciągnięciu wniosków, jakie chciałby wyciągnąć.

  T.R. Aleksandrowicz, Podstawy walki informacyjnej, Warszawa 2016; A. Codevilla, Informing Statecraft. Intelligence for the New Century, New York 1992]

Akcja dezinformacji

Polega zwykle na przekazywaniu konglomeratu informacji, do którego dołączona jest wiadomość fałszywa, kluczowa dla wywołania zakładanego efektu: skłonienia wprowadzonego w błąd przeciwnika/konkurenta do zachowania zgodnego z oczekiwaniami i korzystnego dla dezinformującego.

T.R. Aleksandrowicz, Podstawy walki informacyjnej, Warszawa 2016; A. Codevilla, Informing Statecraft. Intelligence for the New Century, New York 1992]

Ochrona wartości

- **Wolność informacji**: wyrażanie opinii i poglądów bez cenzury jest w Unii Europejskiej gwarantowana przez prawo.
- Dużo czasu zabrało rozpoznanie, że można ją też wykorzystywać do obalania podstawowych zasad UE, a cele pośrednie to uniemożliwienie debaty i krytycznego myślenia.
- Dezinformacja jako skrajna forma nadużycia mediów, zmierzająca do wywarcia wpływu na procesy społeczne i polityczne bywa sponsorowana przez rządy i wykorzystywana w stosunkach międzynarodowych.

Ochrona wartości

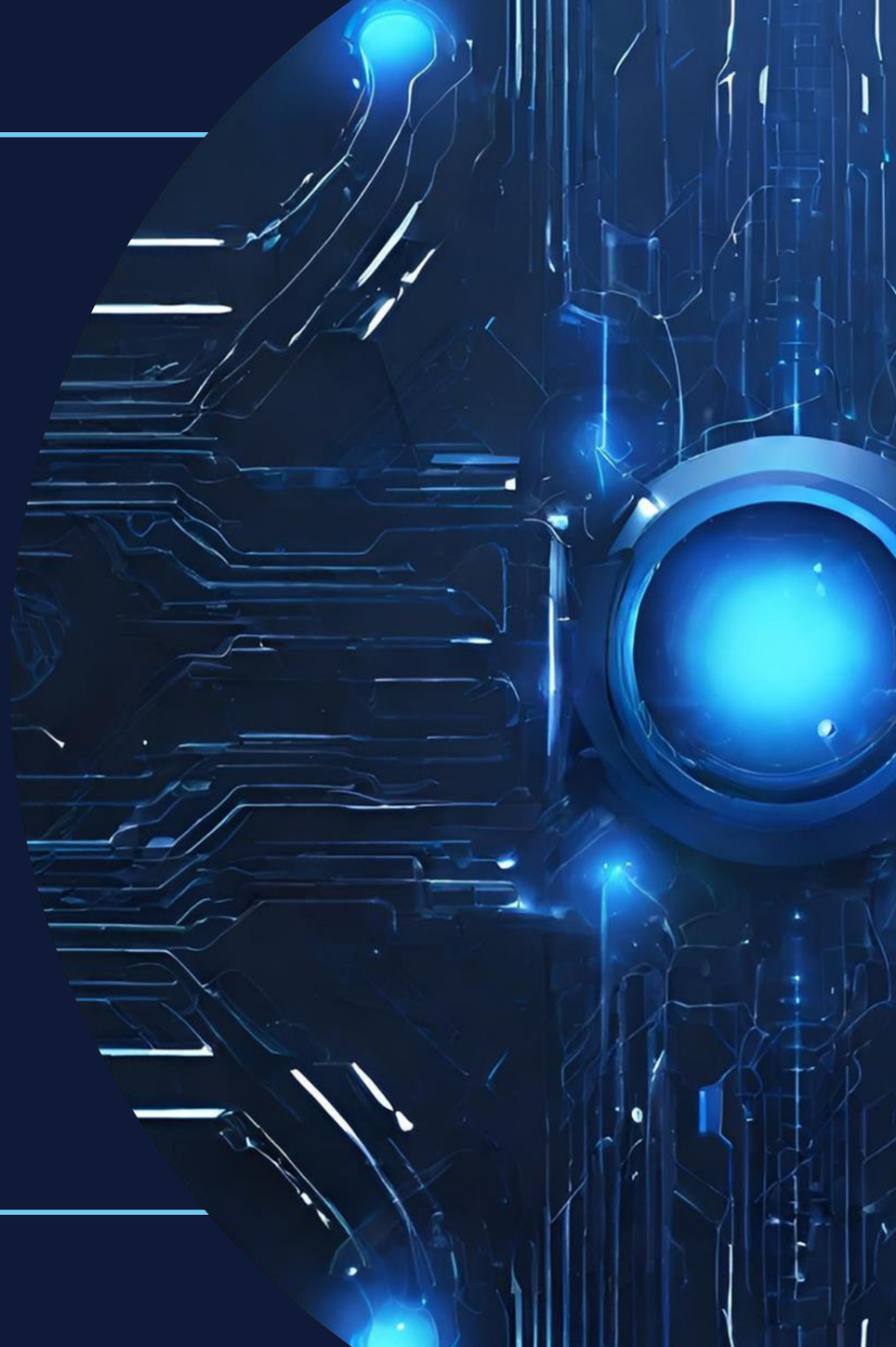
- Dezinformacja umożliwia manipulację opinią publiczną i oddziaływanie w ten sposób z jednej strony na politykę wewnętrzną suwerennych państw, a z drugiej na UE jako całość. Jako niedocenione postrzega EKES (Komitet Rozwoju Ekonomiczno-Społecznego) zwłaszcza ofensywne zdolności cyberoperacji.
- Unijne dokumenty i raporty różnych organizacji przedstawiają liczne dowody rosyjskich kampanii dezinformacyjnych na rzecz Brexit'u, usprawiedliwiania wojny przeciw Ukrainie, ingerencje w wybory

Ochrona wartości

Obok gwarantowanej prawem wolności informacji jako dobra zmniejszającego niepewność, chaos (entropię) niezbędna jest ochrona przed dezinformacją jako złem które entropię zwiększa

2

Jak AI zmienia dezinformację



Rozprzestrzenianie fałszywych wiadomości

- Rozprzestrzenianie fałszywych wiadomości po to, aby przeciwnik lub konkurent zachował się w sposób dający korzyść stosującemu podstęp, jest stosowane od dawna.
- Do XX wieku rozprzestrzenianie fałszywych wiadomości (fake news) przebiegało powolnie mając też ograniczenia przestrzenne i obejmując ograniczoną liczbę ludzi.
- Radykalne znoszenie tych barier przyniósł rozwój Internetu

  i wykorzystanie w nim sztucznej inteligencji.

Sztuczna inteligencja

- Systemy sztucznej inteligencji to systemy maszynowe, zaprojektowane do działania z różnym poziomem autonomii, które mogą po wdrożeniu wykazywać zdolność adaptacji i które – do wyraźnych lub dorozumianych celów – wnioskują, jak generować na podstawie danych wejściowych wyniki, takie jak predykcje, treści, zalecenia lub decyzje, mogące wpływać na środowisko fizyczne lub wirtualne (AI Act)
- Wykorzystanie AI do dezinformacji przejawiało się najpierw generowaniem fałszywych wiadomości tekstowych i głosowych przez boty, a obecnie także tworzeniem deepfakes - zmanipulowanych treści audiowizualnych.

Boty

- Bot [skrót od słowa ‘robot’] – to oprogramowanie stworzone do wykonywania powtarzalnych zadań w bardzo krótkim czasie, oparte na algorytmie sztucznej inteligencji, a jego celem jest naśladowanie ludzkich zachowań i odwzorowywanie ich w internecie.
- Zadaniem dobrych botów jest wyręczanie użytkowników w różnych działaniach, od takich jak zachowanie porządku na forum i usuwanie przekleństw oraz innych niecenzuralnych treści w postach, po pomoc w marketingu i optymalizacji stron internetowych. (np. bot indeksujący, inteligentny asystent, chatbot).

Boty

- Zadaniem złych botów jest wyrządzanie szkód użytkownikom na korzyść cyberprzestępców (np. przez rozprzestrzenianie spamu, pobieranie treści stron, pobierae dane uwierzytelniające, boty zombie)
- Zagrożeniem w sieciach społecznościowych jest manipulacja przez boty podawanymi wiadomościami, jak również brak pełnej świadomości użytkownika o zbieranych o nim informacjach.
- Boty służą również jako narzędzie do zmasowanych ataków (np. na serwisy internetowe) czy rozsyłania zautomatyzowanego spamu.

Boty

- Boty, jako zautomatyzowane konta w mediach społecznościowych, są używane do generowania i rozpowszechniania dużych ilości wiadomości, głównie dla wzmocnienia kontrowersyjnych treści, punktów widzenia.
- Boty polityczne realizują w mediach społecznościowych zadania wzmacniania konkretnego przekazu, masowego rozpowszechniania określonych treści.
- Coraz częściej ma miejsce tworzenie przez boty wrażenia rzeczowej rozmowy oraz sytuacje, w której wiele botów rozmawia ze sobą po to, żeby wciągnąć do tej rozmowy ludzi i przekazać im określone treści, często fake news.

Boty i Botnet

- Botnet to sieć składająca się z komputerów zainfekowanych oprogramowaniem złośliwym (tzw. zombi) wykonujących polecenia i działających na rzecz nieuprawnionej osoby, zwykle hakera. Bot. Paweł Wiszniewski, s.69-70;

por. https://nordvpn.com/pl/blog/bot-co-to-jest/ *Paweł Wiszniewski, Bot [w:]Wielka Encyklopedia Prawa (Tom XXII Prawo Informatyczne. Warszawa 2021, s. 69-70).*

Deep fakes

- Deepfaki to „zmanipulowane treści audiowizualne, przedstawiające ludzi, wydających się robić coś, czego nigdy nie powiedzieli ani nie zrobili, powstałe przy użyciu sztucznej inteligencji, w tym uczenia maszynowego i głębokiego uczenia”.

„Tackling deepfakes in European policy” European Parliament https://www.europarl.europa.eu › STUD › 2021 PDF PE)

- Zagrożenia związane z deepfake'ami mogą mieć charakter psychologiczny, finansowy i społeczny
- Skutki deep fakes mogą obejmować zarówno poziom indywidualny, jak i społeczny.

Strażnicy cyberświata

- Powszechny jest pogląd o potrzebie wzmacniania kontroli nad AI ale sporne kto ją ma sprawować.
- „Sztuczna inteligencja godna zaufania” jest odmiennie pojmowana przez poszczególnych liderów rozwoju systemów AI oraz przez użytkowników i ich przedstawicieli.
- Bezsporne jest, że negatywne skutki rozwoju systemów AI mogą przeważyć nad korzyściami, gdy człowiek straci nad AI kontrolę.

3

Czy jesteśmy odporni na dezinformację?

Sposoby manipulacji

- Przemieszczenie odpowiedzialności
- Przypisywanie winy
- Bierna agresja
- Szantaż emocjonalny
- Posługiwanie się nowomową
- Wciskanie kitu
- Wzbudzanie poczucia krzywdy
- Występowanie w roli ofiary
- Dezinformacja

Sposoby manipulacji

- Narzucanie własnej narracji
- Nieoczekiwana zamiana miejsc
- Pomniejszanie znaczenia problemów
- Wyolbrzymianie znaczenia problemów
- Zastraszanie
- Wprowadzanie podziałów
- Nadmierne uogólnianie
- Czynienie pozornych ustępstw
- Ukrywanie rzeczywistego celu
- Zgłaszanie niewspółmiernych żądań

Odróżnianie prawdy od nieprawdy (OECD)

- Badanie ankietowe przeprowadzone w 2024 r. przez Organizację Współpracy Gospodarczej i Rozwoju (OECD) objęło 42 tys. osób w 21 krajach [w tym w Niemczech i Polsce].
- Respondenci trafnie oznaczali, czy treść była prawdziwa, czy fałszywa, w ok. 60 % przypadków. [najlepiej satyrę - 71 % a najgorzej zmanipulowaną informację - 54 %.
- Nie było rzeczywistej różnicy między zdolnością ludzi do odróżniania faktów od fikcji w sprawach międzynarodowych, zdrowia czy środowiska.

Odróżnianie prawdy od nieprawdy (OECD)

- Ci, którzy ufają mediom społecznościowym jako źródłu informacji, mają mniejszą zdolność wskazywania fałszu (54 proc.) w porównaniu z tymi, którzy w ogóle nie ufają mediom społecznościowym (62 proc.).
- Twierdzenia prawdziwe i fałszywe stworzone przez AI poprawnie rozpoznano w 68 % przypadków a w materiałach generowanych przez ludzi w 60 %.
- Trafność oznaczania prawdziwych oraz fałszywych twierdzeń przez mężczyzn i kobiety była podobna.
- Mężczyźni byli bardziej pewni swojej zdolności do wykrycia dezinformacji w porównaniu z kobietami choć w rzeczywistości nie okazali się lepsi.

Odróżnianie faktów od poglądów i opinii (NASK)

- Ponad połowa polskich internautów stwierdza, że zetknęła się z manipulacją lub dezinformacją, a 35 proc. Polaków z fałszywymi wiadomościami w sieci spotyka się raz w tygodniu lub częściej. Przy tym aż 19 proc. wprost twierdzi, że nie sprawdza wiarygodności internetowych informacji, ani ich źródeł
- Badania dotyczące kompetencji medialnych wykazały, że duża część polskich internautów ma poważne problemy z odróżnianiem opinii od faktu

https://www.nask.pl/pl/aktualnosci/2249,Badania-NASK-ponad-polowa-polskich-internautow-styka-sie-z-manipulacja-i-dezinfo.html

Szacowanie ryzyka – postprawda

- Badania pokazują , że zdolność ludzi do wykrywania fałszu w sieci jest niewielka a ich sytuacja jeszcze się pogarsza, jeśli polegają na mediach społecznościowych jako głównym źródle informacji.
- Rozpowszechnia się „postprawda”, niebezpieczny stan w którym fakty kształtują opinię publiczną w mniejszym stopniu niż emocje i osobiste przekonania. Staje się ona daleko posuniętą relatywizacją faktów, aż do ich zafałszowania.
- Potrzeba lepszych umiejętności korzystania z mediów, w tym wykorzystania dobrych praktyk krajów skandynawskich w zakresie „programów pomagających ludziom rozszyfrować to, co widzą w Internecie”. (kraje skandynawskie, które mają związane z tym długoterminowe projekty, wypadają w badaniach lepiej niż inne kraje).

4

Czy ochrona Instytucjonalna jest dostateczna?

Władze publiczne przeciw deepfakom

- Władze publiczne podejmują różne działania w celu zwalczania dezinformacji. Przykładowo w odniesieniu do wykorzystania deepfake’ów:.
  - W USA w niektórych stanach wprowadzono regulacje penalizujące tworzenie i rozpowszechnianie deepfake’ów o charakterze politycznym i seksualnym.
  - Chiny wymagają oznaczania treści powstałych z użyciem generatywnej AI.
  - Unia Europejska i jej państwa członkowskie zwiększają odpowiedzialność platform internetowych za treści publikowane przez ich użytkowników.

Globalne korporacje sektora IT wobec deepfakes

- Platformy Microsoft i Google wdrażają różnorodne środki, od usuwania zmanipulowanych treści po rozwijanie narzędzi do ich wykrywania:
  - Microsoft opracował narzędzie Video Authenticator do wykrywania deepfake’ów.
  - Google współpracuje z fact-checkerami i usuwa deepfaki mogące powodować szkody.
  - Wielkie korporacje udostępniają bazy danych deepfake’ów do celów badawczych.
  - Google i Microsoft wspierają rozwój technologii watermarkingu.
  - Platforma X (dawniej Twitter) daje dużą swobodę dezinformacji.

Ograniczenia skuteczności

- Rozbieżności między stanem normatywnie postulowanym a realnie obserwowanym są konsekwencją m.in.:
  - konieczności balansowania między ochroną wolności (słowa) i ochroną bezpieczeństwa;
  - konieczność godzenia częściowo sprzecznych postulatów stabilności prawa i nadążania prawa za szybkim rozwojem technologii;
  - wysokich kosztów technologii wykrywania (zwłaszcza deepfake’ów).

5

Jak wzmacniać samoobronę przed dezinformacją?

Zasada ograniczonego zaufania

- „Wracamy do swoistego raju: Nie wiadomo kiedy człowiek może stać się nagi i bosy, gdy da się uwieść swobodzie cyberprzestrzeni” (S. Borkowski).
- Pamiętanie o możliwości ignorancji, nieuczciwości lub złej woli innych użytkowników infostrady.
- Dopuszczanie zawodności zmysłów słuchu i wzroku w rozpoznaniu botów i deepfake’ów.
- Krytyczne myślenie - odfiltrowanie ze złożonej rzeczywistości tego co dla nas istotne i w tym zakresie porównywanie wiadomości z różnych źródeł i sprawdzanie wiarygodności poszczególnych źródeł.

Wiedza o metodach weryfikacji wiadomości

- **Narzędzia techniczne:**
  - AI contra deepfake: https://www.youtube.com/watch?v=zA0W8DadBW0
- **Bazy danych instytucji publicznych i organizacji:**
  - Baza EU vs Disinfo zawiera zestawienie 17963 przypadków dezinformacji wraz ze sprostowaniem: https://euvsdisinfo.eu/pl/; https://euvsdisinfo.eu/de/

Wola i umiejętność sprawdzania i zgłaszania nadużyć w sieci

Treści nielegalne i treści szkodliwe

- **Treści nielegalne** to takie, które naruszają przepisy prawa (najczęściej przepisy kodeksu karnego); m.in.:
    - treści pornograficzne z udziałem małoletniego;
    - treści mogące ułatwić popełnienie przestępstwa o charakterze terrorystycznym.
- **Treści szkodliwe** to materiały, które wywołują u odbiorcy negatywne emocje lub promują niebezpieczne dla zdrowia i życia zachowania; m.in.:
    - Treści prezentujące przemoc lub zachowania agresywne;
    - treści nawołujące do samookaleczeń, samobójstw i innych autodestrukcyjnych zachowań.

Stowarzyszenie INHOPE

- Zgłosić nadużycie w internecie można w narodowych zespołach hotline współdziałających w ramach stowarzyszenia INHOPE (The Association of Internet Hotline Providers).
- Do INHOPE należy ponad 50 zespołów reagujących z całego świata.
- Działania INHOPE są wspierane przez Interpol, Europol, Virtual Task Force, European Financial Coalition, INSAFE, ECPAT oraz globalne firmy sektora informatycznego.

Zgłaszanie treści nielegalnych w Polsce

- Zgłoszenia nielegalnych treści przyjmuje punkt kontaktowy NASK na formularzu pod adresem **https://dyzurnet.pl/** albo na infolinii 0 801 615 005.
    - **Zgłoszenia mogą być anonimowe.**
- Zgłaszać sprawę potencjalnie nielegalnych treści można też na Policję lub do Prokuratury, także gdy pojawia się podejrzenie lub poczucie zagrożenia bez pewności, że doszło do popełnienia przestępstwa.
- Możliwe jest również złożenie zawiadomienia za pomocą https://www.gov.pl/web/gov/zglos-przestepstwo.

Zgłaszanie treści na forach internetowych

- Treści zawierające zniewagę, zniesławienie, obraźliwe komentarze czy wulgaryzmy można też zgłaszać do administratorów strony lub moderatorów forów dyskusyjnych.
  - Najczęściej służą do tego specjalne przyciski lub formularze powiadomień.
  - Jeśli ich nie ma, wykorzystać należy dane zamieszczone w zakładce „Kontakt” danej witryny.

  Rosja wydaje na dezinformację znacznie więcej niż UE na walkę z dezinformacją ale „ludzi dobrej woli jest więcej” i to ich zgłoszenia i aktywność dla wspólnego dobra na portalach społecznościowych zmniejszają siłę rażenia.

Znajomość prawnych instrumentów ochrony i dochodzenia roszczeń

Digital Services Act w Polsce

- W lutym 2024 r. minął termin na wdrażenie do krajowego porządku prawnego Digital Services Act, który nakłada na internetowych usługodawców wiele obowiązków związanych z administrowaniem nielegalnymi treściami.
- Eksperci z zespołów Cyberpolicy i Dyżurnet.pl sprawdzają, jak serwisy internetowe – w tym również hostingodawcy – radzą sobie z moderacją szkodliwych treści zamieszczanych na ich platformach.
    - Wyniki badań oraz wnioski i rekomendacje, które mogą posłużyć jako wskazówki przy opracowywaniu i wdrażaniu efektywnych systemów moderacji zawarto w raporcie dostępnym pod adresem: https://dyzurnet.pl/aktualnosci/wpis/raport-cyberpolicy-i-dyzurnet-pl-od-zgloszenia-do-reakcji

Brońmy uczciwej komunikacji,
a więc pokoju, bo wojna
zaczyna się w słowach

Prawo chroni nas przed destrukcja pokojowej komunikacji

Contact Us

1/3 Wóycickiego St., Warsaw, ZIP 01-938

g.szpor@uksw.edu.pl

https://wpia.uksw.edu.pl/katedry/katedra-prawa-informatycznego/