Cyberhigiena w pracy dziennikarzy

Praktyczne wskazówki

NASK Ministerstwo Cyfryzacji

PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI

1. Bezpieczne hasła i uwierzytelnianie

- Korzystaj z silnych, unikalnych haseł – powinny się składać z co najmniej 14 znaków. Pamiętaj o zasadzie: „jedno hasło – jedno konto".
- Wszędzie tam, gdzie to możliwe, stosuj uwierzytelnianie dwuskładnikowe (2FA). Tzw. drugim składnikiem może być biometria, kod SMS, kod z aplikacji uwierzytelniającej lub najlepiej – klucz bezpieczeństwa.

2. Oprogramowanie i systemy

- Instaluj oprogramowanie z oficjalnych źródeł, np. sklepów z aplikacjami lub stron producentów.
- Regularnie aktualizuj wszystkie urządzenia – zarówno służbowe, jak i prywatne, komputery i smartfony.

3. Kopie zapasowe i dostęp do danych

- Wszystkie kluczowe dane archiwizuj regularnie – najlepiej w chmurze oraz na nośnikach offline. Rozważ stosowanie strategii 3-2-1 (3 kopie danych, 2 różne nośniki, 1 kopia w innej lokalizacji).
- Rozważ szyfrowanie plików i dysków, zwłaszcza w przypadku materiałów wrażliwych lub objętych embargiem.
- Zachowaj ostrożność przy korzystaniu z zewnętrznych nośników danych, np. pendrive'ów.

4. Oddzielanie spraw służbowych od prywatnych

- Oddzielaj przestrzeń zawodową od prywatnej – w ten sposób zmniejszasz ryzyko wycieku danych, przypadkowego ujawnienia informacji czy wykorzystania służbowych zasobów do nieautoryzowanych celów.
- Unikaj logowania się na prywatne konta (np. pocztę, bankowość, media społecznościowe) na urządzeniach służbowych – ograniczy to ryzyko przeniesienia zagrożeń z jednej przestrzeni do drugiej.

5. Ochrona przed phishingiem

- W przypadku e-maili, SMS-ów oraz wiadomości w komunikatorach, zachowaj ostrożność – zwłaszcza jeśli zawierają one linki lub załączniki, a ich treść wzbudza emocje lub wymaga podjęcia natychmiastowych działań – np. pilnej płatności.
- Weryfikuj pełny adres nadawcy e-maili – nie tylko nazwę wyświetlaną w polu „od". Fałszywe adresy często różnią się jedną literą lub pochodzą z nietypowych domen.
- Nie otwieraj załączników do nieoczekiwanych wiadomości pochodzących z niepewnego źródła.

Oprac. na podst. Anna Kwaśnik, *Cyberhigiena w pracy dziennikarza*, w: *Przewodnik po cyberbezpieczeństwie i sztucznej inteligencji dla mediów i twórców cyfrowych.*

Zasady bezpieczeństwa i dobre praktyki w pracy z narzędziami LLM

Praktyczne wskazówki

NASK Ministerstwo Cyfryzacji

PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI

1. Bezpieczeństwo danych osobowych, wrażliwych i niejawnych

- Jeśli nie masz możliwości korzystania z bezpiecznego środowiska (modelu lokalnego lub hostowanego na wewnętrznych serwerach Twojej firmy), nie korzystaj z ogólnie dostępnych czatów opartych na AI w celu przetwarzania danych osobowych, wrażliwych czy informacji niejawnych, o których wyciek możesz się obawiać.
- Przemyśl wykorzystanie anonimizacji danych przed podaniem ich w poleceniu (promp-cie). Możesz użyć fikcyjnych informacji lub zaślepek (np. [nazwisko] lub [kwota]), by wydobyć pożądane odpowiedzi od modelu, a następnie wprowadzić dane faktyczne poza okienkiem czatu.

2. Weryfikacja treści podawanych przez model

- Nie traktuj odpowiedzi modelu jako niepodważalnej wiedzy. Poddaj ją krytycznej analizie. Jeśli informacja nie daje się jednoznacznie zweryfikować, nie ryzykuj jej publikacji.
- Zwracaj szczególną uwagę na podawane przez model cytaty, nazwiska, dane liczbowe czy źródła wiedzy. Pamiętaj, że LLM-y, nawet gdy podają wiarygodne fakty, mogą nie rozumieć występujących między nimi zależności i w związku z tym, wyciągać błędne wnioski na podstawie rzetelnych danych.

3. Ocena naturalności języka

- Zachowaj czujność w zakresie poprawności i naturalności wypowiedzi modelu.
- Sprawdzaj, czy zastosowane związki frazeologiczne występują w słownikach. W przy-padku neologizmów i rzadkich słów upewnij się, że model wybrał właściwą formę fleksyjną, a zapis słowa potwierdzony jest w innych źródłach.

4. Zasady dobrego promptowania

- Jeśli zaczynasz pracę z modelami językowymi lub testujesz nowy model, poświęć czas na sprawdzenie reakcji LLM na interesujące Cię rodzaje zapytań. Z reguły im więcej w poleceniu konkretów, informacji kontekstowych i jasno sformułowanych wytycznych dotyczących stylu czy formatowania, tym bardziej satysfakcjonująca odpowiedź.
- Rozwijaj własny styl w rozmowie z konkretnym modelem językowym, by optymalizować wyniki skrojone pod Twoje potrzeby. Nie zapominaj jednak, że różne modele mogą reagować odmiennie na najczęściej używane przez Ciebie typy zapytań.

5. Świadomość zagrożeń i edukacja

- Dziel się wiedzą i doświadczeniami w zakresie korzystania z LLM-ów w środowisku zawodowym i prywatnym. Im więcej osób będzie świadomych zarówno ich ograniczeń, jak i możliwości, tym większa będzie efektywność i bezpieczeństwo ich wykorzystania.

6. Transparentność

- Oznaczaj treści tworzone lub współtworzone przez AI, by zachować transparentność wobec Twoich odbiorców – zwłaszcza w przypadku materiałów audiowizualnych.

Oprac. na podst. Karolina Seweryn, Anna Kołos, Agnieszka Karlińska, *LLM-y w generowaniu treści: ryzyka i zasady bezpiecznej pracy*, w: *Przewodnik po cyberbezpieczeństwie i sztucznej inteligencji dla mediów i twórców cyfrowych.*

Jak rozpoznać deepfake?

Praktyczne wskazówki

NASK Ministerstwo Cyfryzacji

PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI

Choć jakość deepfejków stale rośnie, większość materiałów nadal zawiera typowe błędy, które mogą zaobserwować uważni odbiorcy. Można wśród nich wyróżnić:

- Zniekształcenia widoczne na krawędziach twarzy lub w miejscach jej zakrywania – na przykład dłonią. Mogą to być podwójne kontury, dodatkowe elementy, nierówne brzegi czy „rozlanie" skóry poza naturalną linię.
- Nienaturalne rozmycia i kontrasty między twarzą a tłem – tzw. efekt maski, gdzie twarz wygląda jak sztucznie nałożona na ciało.
- Ruchy ust opóźnione względem dźwięku lub nieadekwatne do wypowiadanych słów. Mimika nie oddaje emocji lub jest ograniczona do minimalnych ruchów.
- Brak synchronizacji mowy ciała z wypowiedzią – ruchy głowy czy rąk mogą być niespójne z tonem i treścią, np. ktoś mówi emocjonalnie, ale pozostaje całkowicie nieruchomy.
- Nienaturalnie rozmyty lub nadmiernie wyostrzony obszar okolicy ust, zęby przedstawione w zdeformowany sposób.
- Wygładzona tekstura twarzy – nienaturalnie gładka, pozbawiona detali skóra, brak porów, zmarszczek czy przebarwień.
- Dodatkowe lub zniekształcone elementy ciała, np. nadmiar palców, nienaturalne zgięcia dłoni.
- Niespójności szczegółów w tle – powielone przedmioty, krzywe linie, rozmyte napisy, artefakty, błędne odbicia w lustrze lub nieścisłości w oświetleniu i perspektywie.
- Zmieniony akcent, skoki głośności, metaliczny pogłos.
- Maszynowe tempo – przypomina czytanie, z krótkimi odstępami między wyrazami.
- Błędy wymowy – w szczególności nazw własnych lub liczebników, a także niepoprawne ich odmiany. Błędy językowe, tłumaczeniowe i gramatyczne oraz logiczne niespójności.
- Cykliczne, wyraźne oddechy i nienaturalne przerwy.

Jeśli to Ty został_ś ofiarą...

- **Nie pozostawaj biern_!** Nie bój się zgłosić incydentu na policję. Tylko poprzez zwrócenie uwagi organów ścigania na skalę problemu można spodziewać się stworzenia lepszych metod walki z przestępczością internetową. Incydent zgłoś także do zespołu CERT Polska (incydent.cert.pl).
- **Mów o tym, co się stało!** Użyj dostępnych sobie kanałów, by otwarcie oznajmić, że dany materiał nie jest prawdziwy.
- **Skorzystaj z prawa do bycia zapomnianym!** Obywatelom Unii Europejskiej zagwarantowane jest prawo, które pozwala żądać od wyszukiwarek czy portali niezwłocznego usunięcia swoich danych. Warto wziąć to pod uwagę zwłaszcza wtedy, gdy informacje są nieaktualne, nieprawdziwe lub szkodzą Twojej reputacji.

Oprac. na podst. Ewelina Bartuzi-Trokielewicz, Alicja Martinek, *Deepfake: zagrożenia dla środowiska informacyjnego*, w: *Przewodnik po cyberbezpieczeństwie i sztucznej inteligencji dla mediów i twórców cyfrowych.*

Gdzie zgłaszać incydenty?

Praktyczne wskazówki

NASK Ministerstwo Cyfryzacji

PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI

1. Cyberbezpieczeństwo

- Incydenty takie jak złośliwe domeny, podejrzane wiadomości e-mail, deepfejki służące wyłudzaniu danych – zgłaszaj do **CERT Polska** przez stronę **incydent.cert.pl** lub przez aplikację **mObywatel**, w zakładce „Bezpiecznie w sieci”.
- SMS-y z podejrzanymi linkami możesz zgłaszać także przekazując/przesyłając je dalej bezpłatnie **pod numer 8080**.

2. Dezinformacja

- Treści o charakterze dezinformacyjnym, materiały zawierające fałszywe lub zmanipulowane informacje – zgłaszaj do **Ośrodka Analizy Dezinformacji NASK** przez stronę **zglos-dezinformacje.nask.pl**, bezpośrednio na adres mailowy **dezinformacja@nask.pl**.
- Warto również zgłaszać dezinformujące posty bezpośrednio **do administratorów platform** społecznościowych, na których zostały opublikowane.

3. Nielegalne treści

- Materiały przedstawiające seksualne wykorzystywanie dziecka zgłaszaj do zespołu **Dyżurnet.pl** przez stronę **dyzurnet.pl/zglos-nielegalne-tresci**.

4. Deepfejki

- Zgłaszaj wygenerowane materiały **administratorom platform**, na których zostały opublikowane.
- Dodatkowo na adres **deepfake@nask.pl** możesz przesyłać materiały do weryfikacji.

5. Oszustwa

- Zgłaszaj bezpośrednio **na policję**.

Oprac. na podst. *Przewodnik po cyberbezpieczeństwie i sztucznej inteligencji dla mediów i twórców cyfrowych.*