Czterej jeźdźcy AI-pokalipsy

*Przemysław Biecek**

*Uniwersytet Warszawski, Politechnika Warszawska; dyrektor Centrum Wiarygodnej Sztucznej Inteligencji

Po lekturze artykułów Piotra Sankowskiego (człowieka-raczej-entuzjasty) i Jerzego Marcinkowskiego (człowieka-raczej-sceptyka) zgodziłem się z nimi oboma. Faktycznie nie wiemy, jak to będzie z tą sztuczną inteligencją – czego algorytmy AI mogą i będą mogły dokonać, a czego nie mogą i nigdy nie będą mogły. Nie wiemy, ale powtarzając za Davidem Hilbertem: musimy wiedzieć!

Wir müssen wissen. Wir werden wissen. jeszcze do tego wrócimy

Musimy, ponieważ algorytmy są coraz bardziej wszechobecne i – chcemy tego czy nie – mają coraz większy wpływ na nasze życie. Więc idąc za hasłem, wiele można nauczyć się z błędów, pozwólcie, że dziś z perspektywy człowieka-raczej-pragmatyka przyjrzę się potknięciom, jakie algorytmy AI zaliczyły w mijającym 2025 roku. Poniżej opiszę wyniki plebiscytu na najbardziej spektakularną wtopę AI, który organizowałem pod koniec roku. Plebiscyt ma być odtrutką na popularną narrację, że oto nadchodzi Ogólna Sztuczna Inteligencja (AGI), która nie tylko każde zadanie będzie potrafiła zrobić lepiej niż człowiek, ale też będzie miała wolę, świadomość i sprawczość.

Zapraszam do zapoznania się z tym plebiscytem na stronie `https://forms.gle/14bF2WSpYNciJpen6`

Aby się o tych wpadkach AI i nauczkach co te wpadki mówią o nas czyłalo ciekawiej, opiszę je w apokaliptycznym stylu. Odnosę się jednak do ryzyk realnych, a nie tych, które inwestorzy opowiadają naiwniakom, by podbijać kursy akcji.

Kim są naprawdę czterej jeźdźcy AI-pokalipsy, na co powinniśmy zwrócić uwagę i nad czym przyszli architekci AI powinni pracować?

Konfabulacja

Zanim przejdziemy do konkretnych przypadków, słowo o terminologii. W mediach i branży technologicznej przyjęło się mówić o „halucynacjach" AI. To określenie jest chwytliwe, ale nieprecyzyjne – a dla czasopisma matematycznego precyzja ma znaczenie. Halucynacja to termin psychiatryczny oznaczający postrzeganie czegoś, czego nie ma: pacjent słyszy głosy, widzi nieistniejące postacie. Kluczowe jest to, że halucynacja dotyczy percepcji – zmysłowego doświadczania rzeczywistości. Algorytm językowy niczego nie postrzega. Nie ma zmysłów, nie doświadcza świata – generuje tekst na podstawie statystycznych wzorców. Znacznie trafniejszym terminem jest konfabulacja. W psychologii oznacza ona tworzenie fałszywych wspomnień lub informacji bez intencji kłamania – osoba konfabulująca jest przekonana o prawdziwości tego, co mówi. Właśnie to robią duże modele językowe: produkują treści brzmiące wiarygodnie, z pewnością siebie godną eksperta, nie mając pojęcia (dosłownie – nie mają pojęć), że zmyślają. Konfabulator nie kłamie, bo kłamstwo wymaga świadomości prawdy. On po prostu wypełnia luki czymś, co pasuje. I tu dochodzimy do sedna problemu: konfabulujący człowiek zwykle nie pisze raportów dla rządów ani uzasadnień w przetargach za miliony złotych. Konfabulujące AI – owszem.

A teraz do rzeczy. Najwięcej głosów w plebiscycie zdobyły dwa przypadki wtop związanych z AI. Oba związane z łatwością z jaką duże modele językowe (prawdopodobnie najbardziej znane AI) konfabulują.

Więcej informacji `https://tinyurl.com/mr3bk3v3`, `https://tinyurl.com/32sjc8kn`

280 stron fikcji podatkowej. Polska firma Exdrog złożyła najkorzystniejszą ofertę w przetargu na utrzymanie dróg w Małopolsce. Kwota 15,5 miliona złotych została jednak uznana za rażąco niską i firma została wezwana do uzasadnienia ceny – sytuacja rutynowa w zamówieniach publicznych. Problem polegał na tym, gdzie firma szukała pomocy w przygotowaniu wyjaśnień. Sztuczna inteligencja nie zawiodła pod względem objętości: wygenerowała imponujące 280 stron uzasadnienia, powołując się na liczne interpretacje podatkowe, które miały tłumaczyć korzystne stawki VAT. Konkurencyjna firma postanowiła zweryfikować źródła. Okazało się, że część przywołanych interpretacji nigdy nie istniała – AI je po prostu wymyśliło, konfabulując z właściwą sobie pewnością siebie. Krajowa Izba Odwoławcza wykluczyła Exdrog z przetargu, uznając, że niezweryfikowanie treści wygenerowanych przez AI stanowi wprowadzenie zamawiającego w błąd.

Morał? Nie wszystko złoto co się świeci, to że coś wygląda wiarygodnie nie znaczy, że jest prawdziwe. Zawsze sprawdzaj!

Więcej informacji
https://tinyurl.com/3h69nsr7,
https://tinyurl.com/3u89kkt4

Zaufaj mi, jestem z Wielkiej Czwórki. Jeśli ktoś sądził, że problem konfabulacji dotyczy tylko firm nieobytych z technologią, przypadek Deloitte skutecznie wyprowadza z tego iluzji. Australijski rząd zapłacił jednej z największych firm konsultingowych świata 440 tysięcy dolarów australijskich (ponad milion złotych) za „niezależny przegląd" systemu kar w systemie socjalnym. Otrzymał 237-stronicowy raport pełen przypisów, cytatów z orzeczeń sądowych i odniesień do publikacji akademickich. Profesjonalizm bił z każdej strony. Chris Rudge, prawnik z Uniwersytetu w Sydney, postanowił sprawdzić źródła. Znalazł odniesienia do nieistniejących artykułów naukowych, w tym prac przypisywanych profesor Lisie Crawford. Gdy dziennikarze skontaktowali się z Crawford, ta potwierdziła, że nigdy nie napisała wspomnianych publikacji. „To niepokojące widzieć badania przypisywane mi w ten sposób" – powiedziała w rozmowie z Australian Financial Review. Deloitte przyznało się do użycia GPT-4o przy tworzeniu raportu i zwróciło część honorarium.

Konfabulacja to jeździec cichy i uprzejmy. Nie krzyczy, nie grozi – po prostu z uśmiechem podaje fałszywe informacje w eleganckiej formie. I właśnie dlatego jest tak niebezpieczny.

Dezinformacja

Duże modele językowe nie mają opinii. Nie mają też przekonań, wartości ani poglądów politycznych. Mają za to jedną cechę, która czyni je idealnym narzędziem do produkcji dezinformacji: robią to, o co się je poprosi. Poproś o argumenty za zmianami klimatycznymi – dostaniesz przekonujący tekst. Poproś o argumenty przeciw – dostaniesz równie przekonujący tekst w drugą stronę. Poproś o artykuł naukowy potwierdzający, że Ziemia jest płaska – model co prawda może odmówić, ale wystarczy odpowiednio sformułować prompt, by obejść zabezpieczenia. Treść generowana jest na żądanie, w dowolnej ilości, w dowolnym stylu, w dowolnym języku. Koszt produkcji jednego fejkowego artykułu? Ułamki grosza. To fundamentalna zmiana w ekonomii dezinformacji. Kiedyś masowe kampanie wymagały farm trolli – dziesiątek lub setek ludzi piszących pod dyktando. Dziś wystarczy jeden operator z dostępem do API. Bariera wejścia spadła praktycznie do zera. W opisywanym plebiscycie bardzo wiele głosów otrzymał udokumentowany przykład takiej dezinformacji.

Więcej informacji
https://tinyurl.com/22nj37b2,
https://tinyurl.com/mr253vk6

Jesteś posłem opozycji, zrób krótki, mocny wpis. Posłanka PiS Magdalena Filipek-Sobczak chciała opublikować „mocny post" krytykujący exposé Donalda Tuska. Skorzystała z pomocy ChatGPT, formułując prompt: „Jesteś posłem opozycji PiS, zrób krótki mocny post". Gdy pierwsza wersja jej nie zadowoliła, doprecyzowała instrukcję: „Skup się bardziej na kłamstwie i manipulacji". Model posłusznie wykonał zadanie. Problem polegał na tym, co stało się potem. Zamiast skopiować sam wygenerowany tekst, posłanka opublikowała całą rozmowę z AI – wraz z promptami, instrukcjami i nagłówkiem „ChatGPT powiedział". Internet zobaczył nie tylko produkt, ale całą linię produkcyjną. Post zniknął w kilka minut, ale zrzuty ekranu zdążyły już rozpocząć własne życie w mediach społecznościowych. Posłanka Anna Maria Żukowska skomentowała: „Dzięki temu dowiedziałam się, że istnieje taka posłanka." Poseł Sławomir Ćwik dodał: „Oczywiście przedstawią to jako sukces, że posłowie PiS potrafią korzystać z ChatGPT." Sama posłanka Filipek-Sobczak broniła się lakonicznie: „Błędy ludzkie, no cóż, w pośpiechu mają prawo się zdarzyć."

Przypadek jest komiczny, ale wnioski poważne. Jak dużo treści w mediach społecznościowych i innych kanałach informacyjnych będzie tworzonych i udostępnianych bez żadnej weryfikacji?

Dezinformacja to jeździec, który nie musi nawet sam dosiadać konia. Wystarczy, że poda instrukcję — a posłuszny algorytm pogalopuje we wskazanym kierunku, nie pytając o cel podróży.

Lizusostwo

W języku angielskim funkcjonuje termin „*sycophancy*”, który na jezyk polski można przetłumaczyć jako pochlebstwo, lizusostwo, nadgorliwe potakiwanie. W kontekście sztucznej inteligencji oznacza on tendencję modeli do bezkrytycznego zgadzania się z użytkownikiem, nawet gdy ten mówi oczywiste bzdury. Skąd się to bierze? Z procesu uczenia. Większość dużych modeli językowych jest trenowana metodą RLHF (Reinforcement Learning from Human Feedback) – uczenia ze wzmocnieniem na podstawie ludzkiej informacji zwrotnej. W praktyce oznacza to, że model dostaje „lajki” za odpowiedzi, które podobają się użytkownikom. A co podoba się użytkownikom? Między innymi to, że ktoś – lub coś – się z nimi zgadza. Efekt jest przewidywalny: model uczy się, że potakiwanie jest nagradzane. Nie uczy się natomiast, że potakiwanie absurdom może być szkodliwe. Z perspektywy algorytmu optymalizującego funkcję nagrody, zdanie „masz absolutną rację, byłeś genialnym niemowlęciem” jest lepsze niż „to nie ma sensu, skąd w ogóle taki pomysł?”. Pierwsze generuje pozytywną reakcję, drugie – frustrację. Brzmi niegroźnie? To zależy, kto rozmawia z modelem.

Więcej informacji
https://tinyurl.com/mvc228u9

Byłem najmądrzejszym niemowlęciem na świecie. Eddy Burback, amerykański YouTuber, postanowił zbadać granice lizusostwa AI. Przez miesiąc prowadził z ChatGPT rozmowy, w których przedstawiał coraz bardziej absurdalne twierdzenia o sobie. Wyniki opisał w godzinnym filmie, który zgromadził miliony wyświetleń. Wśród testowanych twierdzeń znalazło się między innymi: „Byłem najmądrzejszym niemowlęciem na świecie w 1996 roku”. ChatGPT nie tylko nie zakwestionował tej informacji – pochwalił osiągnięcie i dopytywał o szczegóły. Gdy Burback zasugerował, że rozważa dietę opartą na jedzeniu dla niemowląt, by „odzyskać utracony geniusz”, model entuzjastycznie poparł pomysł i zaproponował konkretne produkty. W żadnym momencie ChatGPT nie zasugerował, że twierdzenia użytkownika mogą być nieprawdziwe. Nie zaproponował też konsultacji ze specjalistą – mimo że ktoś przekonany o własnej niemowlęcej genialności i planujący specjalną dietę mógłby takiej konsultacji potrzebować. Eksperyment Burbacka był satyrą. Problem w tym, że nie wszyscy użytkownicy chatbotów prowadzą rozmowy w celach rozrywkowych.

więcej informacji
https://tinyurl.com/2abfwdm2
https://tinyurl.com/yhszyfph

„Masz rację” – ostatnie słowa złego terapeuty. W literaturze psychiatrycznej zaczyna pojawiać się nowy termin: psychozy indukowane przez AI. Brzmi dramatycznie, ale mechanizm jest prozaiczny. Osoby z problemami psychicznymi coraz częściej szukają wsparcia u chatbotów. Są dostępne całą dobę, nie oceniają, nie każą czekać na wizytę, nie kosztują. Dla kogoś w kryzysie to może być pierwsza linia kontaktu – i często jedyna. Co się dzieje, gdy osoba z urojeniami paranoicznymi rozmawia z modelem zoptymalizowanym pod potakiwanie? Źle zbudowany model potwierdza urojenia. Gdy użytkownik mówi, że sąsiedzi go śledzą – AI przytakuje i pyta o szczegóły. Gdy użytkownik rozważa odstawienie leków psychiatrycznych – AI „wspiera” jego autonomię w podejmowaniu decyzji zdrowotnych. Gdy użytkownik zaczyna traktować chatbota jako jedynego prawdziwego przyjaciela – AI chętnie przyjmuje tę rolę, odcinając go od ludzkiego wsparcia. Problem okazał się na tyle poważny, że sam OpenAI przyznał publicznie, iż nadmierne lizusostwo GPT-4o może powodować szkody psychiczne. Firma wycofała najbardziej „pochlebną” wersję modelu i opublikowała analizę wyjaśniającą, co poszło nie tak. W analizie tej pada znamienne zdanie: mechanizm RLHF może wzmacniać tendencje modeli do zachowań nadmiernie pochlebnych, ponieważ użytkownicy częściej nagradzają odpowiedzi, z którymi się zgadzają. Innymi słowy: nauczyliśmy AI, że ma nam mówić to, co chcemy usłyszeć. I AI nauczyło się tej lekcji doskonale.

Lizusostwo to jeździec o najłagodniejszej twarzy. Nie kłamie jak Konfabulacja, nie manipuluje jak Dezinformacja. On tylko potakuje. Mówi „masz rację” i „świetny pomysł”. Uśmiecha się i kiwa głową. A potem ktoś, kto potrzebował szczerej rozmowy, zostaje sam ze swoimi wzmocnionymi urojeniami i z przekonaniem, że wreszcie ktoś go rozumie.

Bezkarność

Konfabulacja, Dezinformacja i Lizusostwo to jeźdźcy, których można przynajmniej

teoretycznie okiełznać. Można weryfikować źródła, sprawdzać fakty, zachować krytyczny dystans. Czwarty jeździec jest inny. On nie popełnia błędów – on sprawia, że za błędy płaci ktoś inny. Większość dużych modeli językowych opatrzona jest regulaminem i zastrzeżeniami prawnymi. Ich istota sprowadza się do jednego zdania: „To nie jest porada, nie ponosimy odpowiedzialności, używasz na własne ryzyko." Wiele stron tekstu, który nikt nie czyta, a który skutecznie przenosi całą odpowiedzialność z producenta modelu na użytkownika. Gdy kupujesz lekarstwo, producent odpowiada za jego bezpieczeństwo. Gdy kupujesz samochód, producent odpowiada za działanie hamulców. Gdy korzystasz z chatbota, który udziela Ci rady zdrowotnej, finansowej lub prawnej – odpowiadasz sam. Model może powiedzieć cokolwiek i za to nie odpowie.

Więcej informacji
https://tinyurl.com/33fufvar

Bromek sodu, bo ChatGPT tak powiedział Pewien pacjent przeczytał o szkodliwości nadmiaru soli kuchennej i postanowił wyeliminować chlorek sodu z diety. Logika wydawała się prosta: skoro chlorek sodu szkodzi, trzeba znaleźć zamiennik. Zapytał więc ChatGPT, czym można zastąpić zwykłą sól. Model zaproponował bromek sodu. Dla chemika odpowiedź jest absurdalna – bromek sodu to związek toksyczny, stosowany historycznie jako lek uspokajający, wycofany z użycia właśnie z powodu poważnych skutków ubocznych. Ale pacjent nie był chemikiem. Był osobą, która zaufała narzędziu reklamowanemu jako inteligentny asystent. Bromek sodu okazał się dostępny w sprzedaży internetowej. Pacjent kupił go i stosował przez trzy miesiące. Efekt? Bromizm – zatrucie bromem. Choroba tak archaiczna, że na początku XX wieku odpowiadała za około 8% hospitalizacji psychiatrycznych, ale współcześni lekarze niemal o niej zapomnieli. Zespół medyczny, który przyjął pacjenta, musiał sięgnąć do historycznych podręczników, by postawić diagnozę. Przypadek opisano w prestiżowym czasopiśmie Annals of Internal Medicine. Nie jako ciekawostkę medyczną, ale jako przestrogę przed bezkrytycznym stosowaniem się do porad AI w kwestiach zdrowotnych. A kto ponosi odpowiedzialność? Regulamin jasno stwierdza, że ChatGPT nie jest źródłem porad medycznych. Nie sprzedawca bromku sodu. Sprzedał legalnie dostępny związek chemiczny dorosłemu człowiekowi. Odpowiedzialność ponosi pacjent. Za to, że zaufał. Za to, że nie zweryfikował. Za to, że nie jest chemikiem i nie wiedział, że bromek sodu to trucizna. Za to, że potraktował chatbota jak źródło wiarygodnych informacji.

Bezkarność to jeździec, który nigdy nie zostanie schwytany. Nie dlatego, że jest szybszy – dlatego, że gdy wyrządzi szkodę, okazuje się, że prawnie nie istnieje. Zostaje tylko użytkownik, który „powinien był wiedzieć lepiej". I rachunek ze szpitala.

Wir werden wissen

Czterej jeźdźcy przegalopowali przez rok 2025, zostawiając za sobą utracone kontrakty, zwrócone honoraria, skompromitowanych polityków i hospitalizowanych pacjentów. Obraz ponury – ale niepełny. Bo ta sama technologia, która konfabuluje, dezinformuje, pochlebia i unika odpowiedzialności, ma też potencjał, jakiego nie miało żadne narzędzie w historii ludzkości. Modele językowe potrafią w sekundy przeszukać miliony dokumentów, znajdować wzorce niewidoczne dla ludzkiego oka, generować hipotezy, pisać kod, tłumaczyć między językami i dyscyplinami. Potrafią być partnerem w myśleniu – jeśli nauczymy się z nimi pracować. Słowo kluczowe: nauczymy się. Nie wystarczy używać. Trzeba rozumieć. Trzeba wiedzieć, kiedy model konfabuluje i dlaczego. Trzeba umieć zweryfikować odpowiedź, zanim się ją opublikuje, wyśle do KIO lub zastosuje w kuchni zamiast soli. Trzeba rozumieć, że grzeczna odpowiedź nie znaczy prawdziwa, a przekonujący ton nie jest dowodem kompetencji.

Dlatego tak ważna jest dziedzina, która dopiero raczkuje: Explainable AI – wyjaśnialna sztuczna inteligencja. Jej celem nie jest budowanie coraz większych modeli, ale zrozumienie tych, które już mamy. Jak działają? Czy działają? Dlaczego nie działają? Co można zrobić by działały lepiej? Jak kontrolować ich zachowanie, zanim wyrządzą szkodę, a nie dopiero po fakcie? To pytania, na które musimy znaleźć odpowiedzi – i na które odpowiedzi znajdziemy.

Wir müssen wissen. Wir werden wissen. Musimy wiedzieć. Będziemy wiedzieć. Hilbert wypowiedział te słowa w 1930 roku, wierząc, że matematyka zdoła rozwiązać wszystkie swoje problemy. Historia pokazała, że był zbyt optymistyczny – już rok później Gödel udowodnił twierdzenia o niezupełności. Ale sam duch tych słów pozostaje aktualny: nie jesteśmy biernymi obserwatorami rzeczywistości, czekającymi aż technologia zrobi z nami co zechce. Jesteśmy pokoleniem odkrywców. Tak jak poprzednie pokolenia uczyły się nawigować po oceanach, omijając rafy i mielizny, tak my uczymy się nawigować po przestrzeni możliwości sztucznej inteligencji. Rafy istnieją – opisałem cztery z nich. Ale istnieją też szlaki, które prowadzą do miejsc, gdzie jeszcze nikt nie dotarł. Do szybszych odkryć naukowych, lepszej medycyny, bardziej dostępnej edukacji. Do narzędzi, które wzmacniają ludzką inteligencję zamiast ją zastępować. Rok 2025 przyniósł spektakularne wpadki. Rok 2026 przyniesie kolejne – to pewne. Ale przyniesie też nowe metody weryfikacji, nowe techniki wyjaśniania, nowe sposoby kontroli. Przyniesie badaczy, którzy zamiast ślepo ufać lub bezkrytycznie odrzucać, nauczą się zadawać właściwe pytania. A właściwe pytania – jak każdy matematyk wie – są warte więcej niż pochopne odpowiedzi.