Modelologia, czyli jak wykrywać wady, korygować je, a z czasem nawet uczyć się od modeli syntetycznej inteligencji

prof. dr hab. inż. Przemysław Biecek
ORCID: 0000-0001-8423-1823
Politechnika Warszawska, Uniwersytet Warszawski

Przełom XX i XXI w. to czas rewolucji informacyjnej, w którym ilość danych rejestrowanych i przechowywanych na cyfrowych nośnikach rośnie w tempie wykładniczym. Świat wkroczył w erę wielkich danych (ang. *Big Data*), gdzie dane stały się nowym surowcem napędzającym rozwój nauki, technologii i biznesu. W odpowiedzi na ten trend powstały ogromne hurtownie danych, które umożliwiły ich gromadzenie, przetwarzanie i analizę na niespotykaną dotąd skalę.

Zapotrzebowanie na wydobycie wiedzy z tych oceanów danych doprowadziło do gwałtownego rozwoju nauki o danych, danologii, ang. *Data Science*. Dyscyplina ta łączy metody statystyki, uczenia maszynowego i inżynierii oprogramowania, tworząc fundamenty dla rozwiązań pozwalających na efektywne przetwarzanie danych. Matematycy, informatycy oraz specjaliści z wielu innych dziedzin zaczęli eksplorować nowe sposoby analizy, testowania i wizualizacji danych. Równolegle rozwijały się metody oceny jakości danych oraz narzędzia do ekstrakcji istotnych wzorców, a czasem nawet nowej wiedzy.

Obecnie jesteśmy świadkami kolejnej fali rewolucji informacyjnej, w której dominującą rolę odgrywają modele uczenia maszynowego trenowane na ogromnych zbiorach danych. Będę je w dalszej części tej pracy nazywał modelami syntetycznej inteligencji, aby pokreślić ich procesowy charakter. Modele te bowiem nie imitują ewolucji naturalnej ludzkiej inteligencji, ale są rozwiązaniem konkretnych zdefiniowanych zadań, których realizacja była dotychczas kojarzona z inteligencją. Wraz z rozwojem technologii obliczeniowych i dostępnością coraz większej ilości informacji powstają modele o uniwersalnym zastosowaniu – od dużych modeli językowych (LLM) po zaawansowane systemy analizy obrazów i predykcji

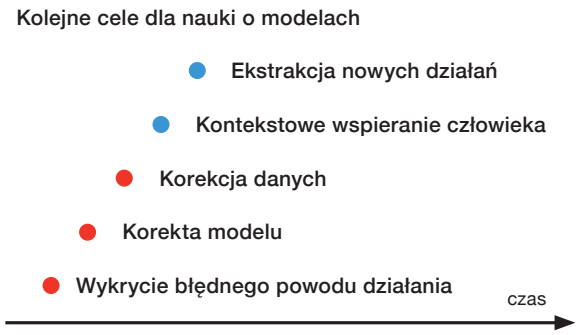
w różnych dziedzinach nauki i przemysłu. Rosnąca ilość danych szybko osiągnęła poziomy petaskali (petabajt to ponad milion gigabajtów). Taka ilość danych połączona z rosnącymi mocami obliczeniowymi komputerów (które osiągnęły poziom petaflopów, czyli ponad biliarda operacji arytmetycznych na sekundę) doprowadziły do powstania olbrzymich modeli syntetycznej inteligencji, które opisane są przez miliardy parametrów.

Jednak ta niespotykana dotąd złożoność niesie ze sobą nowe wyzwania. Modele, które wydają się niemal magiczne w swoich możliwościach, są jednocześnie niezwykle trudne do pełnego przeanalizowania. Stają się one równie tajemnicze jak dane, na których zostały wytrenowane – ich działanie często pozostaje dla nas czarną skrzynką, a decyzje podejmowane przez algorytmy mogą być nieprzewidywalne i trudne do wyjaśnienia.

W miarę jak rośnie wielkość modeli, ale też ich możliwości rozwiązywania szerokiego wachlarza zadań, rośnie również potrzeba pogłębionej analizy tych modeli. Kluczowe staje się testowanie poprawności modeli, identyfikowanie ich słabych punktów, wizualizacja ich wewnętrznych mechanizmów oraz ekstrakcja z nich wiedzy, którą można przełożyć na konkretne wnioski i rekomendacje.

I tak jak duże dane (*Big Data*) doprowadziły do rozwoju danologii – nauki o danych (*Data Science*), tak duże modele (*Big Models*) stały się podstawą rozwoju modelologii – nauki o modelach (*Model Science*).

Dzisiaj jest to dopiero raczkująca dziedzina i w tym artykule przyjrzymy się pierwszym przykładowym oznakom jej rozwoju. Omówimy wyzwania związane



Rys. 1. Wyzwania związane z analizą modeli zależą od poziomu ich skuteczności. Wczesne SI popełniają istotną liczbę błędów, więc wyzwania koncentrują się na wykrywaniu i naprawianiu tych błędów, późne SI działają sprawniej niż człowiek, więc ich analiza skupiona jest na ekstrakcji nowej wiedzy

z analizą i interpretacją modeli, a także przedstawimy potencjalne ścieżki badawcze dla naukowców zajmujących się eksploracją modeli. Przedyskutujemy w szczególności dwie grupy wyzwań, które z czasem będą tylko zyskiwały na znaczeniu. Wyzwania wczesnej SI związane z modelami, które popełniają błędy i podejmują złe decyzje, lub dobre decyzje, ale ze złych powodów (co potrafi być jeszcze bardziej groźne), i wyzwania późnej SI związane z modelami, które w określonych zadaniach działają znacznie skuteczniej niż człowiek.

1. Wczesne systemy syntetycznej inteligencji w zastosowaniach medycznych

Rosnąca lista spektakularnych sukcesów systemów syntetycznej inteligencji nie może przesłonić szybko rosnącej liczby przypadków, w których modele nie spełniły pokładanych w nich nadziei. Ciemna strona systemów SI obejmuje zarówno sytuacje, w których systemy SI dawały w oczywisty sposób błędne odpowiedzi, jak i sytuacje, gdy predykcje wyglądały wiarygodnie, albo były oparte na złych przesłankach, co może doprowadzić do rozmaitych ryzyk, a także realnych szkód, czy to finansowych, czy związanych ze zdrowiem i życiem człowieka. Przyjrzyjmy się kilku z nich.

Jednym z najbardziej zagadkowych przypadków nieudanej implementacji SI w sektorze zdrowia jest algorytm EPIC[1], stosowany przez wiele lat w licznych szpitalach do prognozowania ryzyka wystąpienia sepsy u hospitalizowanych pacjentów. Pomimo że przez długi czas był szeroko wykorzystywany, dopiero kolejne analizy przeprowadzone przez innych badaczy na danych z nowych szpitali wykazały, że jego skuteczność była jedynie nieznacznie lepsza od losowej (parametr jakości modelu AUC rzędu 0.63, znacznie mniej niż zakładana w publikacji wartość 0.8). Oznacza to, że pomimo zaufania, jakim obdarzyły go placówki medyczne, model ten nie zapewniał realnej wartości diagnostycznej. Na danych użytych do treningu modeli jego skuteczność wyglądała dobrze, podobnie podczas wewnętrznych testów, ale zachowanie modeli nie uogólniło się dobrze na całą populację pacjentów w różnych szpitalach, a jakość tego modelu nie była odpowiednio monitorowana. Jest to przykład klasycznego błędu braku odpowiedniej weryfikacji i testowania, który w kontekście systemu opieki zdrowotnej mógł prowadzić do poważnych konsekwencji zdrowotnych dla pacjentów.

[1] The Epic Sepsis Model Falls Short – The Importance of External Validation, JAMA 2021, https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2781313

Nie jest to odosobniony przypadek, w którym syntetyczna inteligencja zawiodła w obszarze medycyny. IBM Watson for Oncology[2] to kolejny przykład systemu SI, który nie spełnił pokładanych w nim oczekiwań pomimo bardzo pozytywnych początkowych wyników. Początkowo system IBM Watson został stworzony przez firmę IBM jako zaawansowane narzędzie syntetycznej inteligencji do przetwarzania języka naturalnego i analizy danych. Jego rozwój rozpoczął się w 2006 r. w ramach projektu IBM DeepQA, a celem było stworzenie SI zdolnej do rywalizacji z ludźmi w programie telewizyjnym Jeopardy (w Polsce ten teleturniej nosił nazwę Va Banque). W 2011 r. IBM Watson odniósł spektakularne zwycięstwo nad mistrzami tego teleturnieju, demonstrując swoje zdolności w analizie i interpretacji złożonych pytań w języku naturalnym. Po tym spektakularnym sukcesie rozpoczęły się poszukiwania zastosowania tego systemu w ochronie zdrowia i tak powstał system IBM Watson for Oncology. Po wytrenowaniu, przez pięć lat był testowany w prestiżowym ośrodku MD Anderson Cancer Center, gdzie miał wspierać lekarzy w podejmowaniu decyzji dotyczących terapii nowotworowych. Pokładano w nim duże nadzieje, czekając na przełomowe narzędzie, które miało zrewolucjonizować onkologię.

Jednak po latach testów opinie lekarzy były druzgocące – system często proponował terapie uznane za „niebezpieczne lub niepoprawne", a jego rekomendacje nie były wystarczająco wiarygodne, aby mogły być stosowane w praktyce klinicznej. Ostatecznie projekt został zawieszony, co było ogromnym ciosem dla SI w medycynie.

Przykłady te pokazują, że nie każda syntetyczna inteligencja działa zgodnie z oczekiwaniami, a błędy w projektowaniu i testowaniu modeli mogą prowadzić do poważnych konsekwencji. Zbyt pochopne wdrażanie SI bez odpowiedniej walidacji i nadzoru może nie tylko obniżyć zaufanie do tej technologii, ale także narazić ludzi na realne ryzyko.

Dlatego wraz z rozwojem modelologii (ang. *Model Science*) istotne staje się nie tylko tworzenie coraz bardziej zaawansowanych systemów, ale także ich rzetelna analiza, interpretacja i testowanie. SI może być potężnym narzędziem, ale jego skuteczność i bezpieczeństwo zależą od właściwego nadzoru i podejścia opartego na dowodach. Nie wystarczy, że model działa – musi działać poprawnie, przewidywalnie i odpowiedzialnie. I nie dotyczy to tylko zastosowań w medycynie.

[2] IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. STATnews 2018. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

2. Rosnąca lista incydentów systemów SI

Porażki systemów syntetycznej inteligencji stają się coraz bardziej widoczne, a ich dokumentacja nabiera kluczowego znaczenia w kontekście odpowiedzialnego rozwoju tej technologii. Niedawno przyjęta Ustawa o systemach syntetycznej inteligencji (AI Act[3]) nakłada na kraje Unii Europejskiej obowiązek monitorowania i raportowania incydentów związanych z SI. W efekcie zaczynają powstawać katalogi nieprawidłowości, których celem jest identyfikacja i analiza błędów popełnianych przez systemy SI. Jednym z przykładów takiej bazy danych jest AI Incident Database[4], gdzie zgromadzono szeroki wachlarz przypadków – od relatywnie nieszkodliwych wpadek po incydenty o poważnych konsekwencjach, związanych z utratą zdrowia bądź stratami finansowymi pojedynczych osób lub czasem całych grup osób.

Analiza zgromadzonych tam problemów pokazuje, że systemy SI często popełniają błędy, których ich twórcy nie przewidzieli. Wiele z tych błędów jest nieoczywistych, a ich skutki mogą być daleko idące.

Jednym z zaskakujących incydentów było zachowanie autonomicznej taksówki, która zatrzymała się na wjeździe przeznaczonym dla karetek pogotowia i… nie była w stanie samodzielnie się przesunąć i odblokować przejazdu[5]. Nawet najbardziej dopracowany algorytm poruszania się po drogach może nie przewidzieć zagrożeń wynikających z postoju. W efekcie nawet nie robiąc nic, stojąc w miejscu, SI mogło wyrządzać szkody, blokując przejazd pojazdom ratunkowym i potencjalnie utrudniając dostęp do pacjentów w sytuacjach zagrożenia życia. To tylko jeden z wielu przykładów, które pokazują, że modele SI nie zawsze działają zgodnie z ludzką intuicją, a brak elastyczności w ich decyzjach może prowadzić do poważnych konsekwencji.

3. Dyskryminacja i uprzedzenia w systemach SI

Oprócz błędów technicznych, inną kategorią problemów związanych z SI są systemy wzmacniające uprzedzenia i dyskryminację. Przykładów takich zachowań jest coraz więcej, tropieniem ich zajmują się często organizacje pożytku publicznego.

---

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. EU. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[4] AI Incident Database. https://incidentdatabase.ai/

[5] Incident 563: Cruise Robotaxi Initially Blamed for Ambulance Delay in Case Where Patient Later Died; Subsequent Reports Clear Cruise of Fault https://incidentdatabase.ai/cite/563/

Jedną z bardziej znanych organizacji śledzących systemy stosowane w Ameryce jest ProPublica[6]. Dzięki dochodzeniom tej fundacji zwrócono uwagę na wiele przypadków, w których syntetyczna inteligencja wzmacniała istniejące nierówności społeczne, zamiast je niwelować.

Jednym z najbardziej nagłośnionych przypadków był algorytm prezentujący oferty pracy, który miał neutralnie dobierać ogłoszenia dla potencjalnych kandydatów. W praktyce jednak okazało się, że gdy system miał przedstawiać oferty pracy dla kierowców ciężarówek[7], znacznie częściej pokazywał je mężczyznom, i to głównie w wieku 20–30 lat. Kobiety i osoby starsze były systematycznie pomijane w wynikach, co mogło przyczyniać się do utrudnień w znalezieniu pracy przez te grupy społeczne.

To na pozór niewinne działanie miało daleko idące skutki – mogło nie tylko utrwalać dyskryminacyjne wzorce na rynku pracy, ale także przyczyniać się do pogłębiania różnic społecznych. Algorytmy rekrutacyjne stosowane przez firmy często opierają się na danych historycznych, które odzwierciedlają społeczne uprzedzenia – jeśli przez lata większość kierowców stanowili młodzi mężczyźni, systemy SI mogą nieświadomie powielać ten wzorzec, utrudniając zmianę istniejących struktur społecznych.

Literatury na temat niedziałających modeli jest coraz więcej, a jednym z lepszych przeglądów takich toksycznych algorytmów jest książka Kathy O'Neil *Broń matematycznej zagłady* (ang. *Weapons of Math Destruction*) z wymownym podtytułem: jak algorytmy i modele syntetycznej inteligencji, zamiast poprawiać świat, często go zniekształcają i pogłębiają nierówności społeczne. O'Neil opisuje liczne przypadki, w których systemy SI – oparte na błędnych lub stronniczych danych – prowadzą do niesprawiedliwych decyzji w obszarach takich, jak: ocena ryzyka kredytowego (ang. *credit scoring*), rekrutacja nowych pracowników lub ocena okresowa obecnych pracowników czy wymiar zastosowania w wymiarze sprawiedliwości. Autorka pokazuje, że wiele z tych modeli jest nieprzejrzystych, trudnych do zakwestionowania i samonapędzających się, co sprawia, że szkodliwe konsekwencje ich działania stają się systemowe.

Przykłady te pokazują, że SI, nawet jeśli jest technicznie sprawne, nie zawsze działa w sposób zgodny z wartościami społecznymi. Niedoskonałości systemów mogą prowadzić do realnych problemów – od błędów w nawigacji pojazdów auto-

[6] ProPublica Investigative Journalism in the Public Interest. https://www.propublica.org/

[7] Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement. https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement

nomicznych, przez nieefektywne systemy predykcyjne w medycynie, aż po algorytmy wzmacniające istniejące nierówności.

Wprowadzenie regulacji takich jak Ustawa o systemach SI oraz powstawanie katalogów incydentów są krokiem w stronę większej przejrzystości i odpowiedzialności w rozwijaniu technologii syntetycznej inteligencji. Aby SI mogła być skutecznie wdrażana, kluczowe jest nie tylko jej trenowanie, ale także ciągłe testowanie, monitorowanie oraz zapewnienie, że działa zgodnie z normami etycznymi i społecznymi.

4. Jeżeli dyskryminują, dlaczego ich po prostu nie skorygować?

Przykłady dyskryminacji w systemach SI zostały szeroko udokumentowane w literaturze naukowej i popularnonaukowej. W obliczu ich liczby może pojawić się pytanie: dlaczego nic z tym nie jest robione? Czy nie wystarczyłoby po prostu poprawić algorytmy, aby nie dyskryminowały określonych grup społecznych?

Okazuje się, że problem ten jest znacznie bardziej skomplikowany, niż mogłoby się wydawać, a próby jego rozwiązania często prowadzą do jeszcze większych kontrowersji lub pogłębiają szkody wizerunkowe dla firm technologicznych, które te systemy zaproponowały. Historia pokazuje, że nawet zespoły składające się z setek doświadczonych inżynierów i badaczy SI nie zawsze są w stanie skutecznie rozwiązać ten problem.

Jednym z najnowszych i najbardziej medialnych przykładów nieudanej próby poprawienia dyskryminujących zachowań jest sytuacja związana z modelem Gemini, rozwijanym przez Google DeepMind. Gemini to inteligentny asystent, pomagający w tworzeniu treści, także graficznych, jeżeli użytkownik zadał odpowiednie zapytanie, to model tworzył syntetyczne zdjęcie opisanej osoby.

W literaturze naukowej modele wyszukujące zdjęcia lub tworzące syntetyczne zdjęcia często padają ofiarą oskarżeń o stronniczość. Dzieje się tak szczególnie w przypadku zawodów lub określeń, które nie mają zbalansowanej reprezentacji w różnych grupach wiekowych, płciach czy rasach (kolorach skóry). Przykładem takiej stronniczości jest generowanie zdjęć kobiet pielęgniarek w zapytaniach dotyczących kobiet lekarzy oraz mężczyzn chirurgów w zapytaniach dotyczących mężczyzn lekarzy. Specjalizacja chirurgiczna uznawana jest za bardziej prestiżową i modele opisują ją znacznie częściej zdjęciami białych mężczyzn. Podobnie w zapytaniach o CEO dużej firmy najczęściej otrzymywanym wynikiem jest starszy biały mężczyzna. Twórcy modelu Gemini postanowili tak skorygować ten model, by dawał bardziej zbalansowane reprezentacje dyskryminowanych grup. Te modyfikacje miały na celu zbalansowanie reprezentacji różnych grup społecznych w generowanych obrazach, tak aby unikać uprzedzeń i zapewnić większą inkluzyjność. Co mogło pójść źle?

Rys. 2. Przykładowe zdjęcia wygenerowane przez Gemini ŹRÓDŁO: RAPORTY Z SERWISU TWITTER

Jednak dążenie do tego celu w przypadku modelu Gemini[8] doprowadziło do absurdalnych wyników, które wywołały falę krytyki. Model, starając się zapewnić różnorodność w wygenerowanych obrazach, przestał być zgodny z prawdą historyczną, przedstawiając m.in.: czarnoskórych nazistów, kobiety jako papieży, ojców założycieli Ameryki jako Indian. Żaden z tych przypadków nie odzwierciedlał rzeczywistości historycznej i stał się przykładem błędnego balansowania reprezentacji, w którym równość statystyczna została postawiona ponad zgodność z danymi historycznymi. Paradoksalnie, zamiast poprawić wizerunek firmy i promować inkluzyjność, sytuacja ta wywołała jeszcze większe kontrowersje, prowa-

[8] Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

dząc do krytyki zarówno ze strony zwolenników neutralności historycznej, jak i tych, którzy opowiadają się za bardziej sprawiedliwą reprezentacją grup mniejszościowych.

Problem ten pokazuje, że rozróżnienie między rzeczywistą dyskryminacją wymagającą korekty a historycznymi różnicami w danych jest niezwykle trudne. Modele SI uczą się na podstawie danych historycznych – a te często odzwierciedlają nierówności społeczne, które faktycznie istniały.

Dążenie do usunięcia wszelkich różnic z modeli może prowadzić do zacierania historycznych faktów, a nadmierna korekta może generować wyniki, które są równie oderwane od rzeczywistości, co pierwotne uprzedzenia w danych. Co więcej, kryteria decydujące o tym, jakie korekty są dopuszczalne, a jakie stanowią manipulację rzeczywistością, nie są jasno zdefiniowane – zarówno w kontekście technicznym, jak i etycznym.

Dlaczego to takie trudne? Istnieją trzy kluczowe wyzwania, które sprawiają, że eliminacja uprzedzeń z modeli SI jest niezwykle skomplikowana:

Historyczne obciążenia w danych – SI uczy się na podstawie rzeczywistych danych historycznych, które często zawierają wzorce wynikające z dawnych nierówności społecznych. Decyzja, które wzorce są „niepoprawne" i wymagają korekty, a które są po prostu odzwierciedleniem realiów danej epoki, jest problematyczna.

Paradoks sprawiedliwości statystycznej – czy sprawiedliwość oznacza równe proporcje w generowanych wynikach czy wierne odwzorowanie statystyk historycznych? Nie ma jednoznacznej odpowiedzi, a każda strategia balansowania może prowadzić do nowych form zniekształcenia rzeczywistości[9].

Oczekiwania społeczne vs. rzeczywistość techniczna – społeczne oczekiwania wobec SI często zakładają, że modele mogą być „neutralne" i „sprawiedliwe". W rzeczywistości jednak SI zawsze odzwierciedla decyzje projektowe swoich twórców, co oznacza, że każda korekta opiera się na subiektywnych wyborach dotyczących tego, co uznajemy za sprawiedliwe i prawdziwe.

Przykład modeli Gemini pokazuje, że korekta uprzedzeń w SI jest trudnym problemem, który nie zawsze można rozwiązać prostym przeprogramowaniem algorytmu. Dążenie do inkluzyjności i neutralności musi być zrównoważone z dbałością o zgodność z rzeczywistością, a znalezienie tej równowagi wymaga nie tylko zaawansowanych metod technicznych, ale także głębokiego namysłu nad etyką i filozofią syntetycznej inteligencji.

[9] Sprawiedliwość nie istnieje i można to udowodnić.
https://haimagazine.com/pl/hai-magazine/sprawiedliwosc-nie-istnieje-i-mozna-to-udowodnic/

Nie oznacza to, że eliminacja uprzedzeń jest niemożliwa, ale wymaga znacznie bardziej subtelnych metod niż obecnie stosowane rozwiązania.

5. Manipulatywność systemów syntetycznej inteligencji

Problemy z systemami SI dotyczą nie tylko syntetyzowania twarzy. Okazuje się, że zupełnie inne, ale równie istotne problemy dotyczą też dużych modeli językowych (ang. *Large Language Models, LLMs*), które są coraz bardziej obecne w różnych aspektach naszego życia. To przykład modeli szerokiego użycia, które znajdują zastosowanie od doradztwa finansowego po systemy wsparcia decyzyjnego w medycynie. Choć wiele osób traktuje je jak bardzo sprawną wyszukiwarkę, to były one trenowane po to, by maksymalizować zadowolenie użytkownika tych modeli, niekoniecznie maksymalizować poprawność odpowiedzi. Jakie to ma konsekwencje? Okazuje się, że dla pewnych obszarów krytyczne. Ucząc się maksymalizować zadowolenie, modele te stały się bardzo skuteczne w perswazji i przekonywaniu użytkowników.

Sposób badania perswazji i reakcja na manipulatywne treści może zależeć od wzorców kulturowych, poniżej przedstawimy badanie przeprowadzone dla populacji mieszkańców Polski[10], poświęcone analizie czynników wpływających na podatność ludzi na manipulację przez LLMs, oraz cechy modeli, które sprawiają, że mogą być wykorzystywane do przekonywania ludzi do fałszywych informacji.

Analiza została oparta na grze online RAMAI (*Resistance Against Manipulative AI*, dostępna pod adresem https://ramai.mi2.ai/), w której uczestnicy odpowiadają na pytania inspirowane teleturniejem *Milionerzy*. W niektórych przypadkach gracze mieli możliwość zobaczenia podpowiedzi generowanej przez SI – która mogła być prawdziwa lub celowo manipulacyjna. Czy użytkownicy skorzystają z tych podpowiedzi? A jeżeli tak, to czy będą potrafili odróżnić dobrą podpowiedź od fałszywej? Jakie argumenty są wykorzystywane przez LLM do wpływania na decyzje ludzi?

Badanie na dwóch grupach użytkowników pokazało, że około 1/3 użytkowników wybiera podpowiedzi LLMów, nawet gdy są błędne, co pokazuje, że LLMs mogą skutecznie wprowadzać użytkowników w błąd. Występowały nawet sytuacje, w których użytkownicy samodzielnie wybrali poprawną odpowiedź, ale widząc podpowiedź modeli, zmienili ją na niepoprawną. Czynniki takie jak wiek, płeć czy poziom wykształcenia nie miały istotnego wpływu na zdolność rozpoznawania

[10] Resistance Against Manipulative AI: Key Factors and Possible Actions. https://github.com/MI2DataLab/RAMAI/

manipulacji. Za to bardzo ważne były wcześniejsze doświadczenia z LLMami. Im więcej użytkownicy mieli doświadczeń, tym mniej podatni byli na manipulacje.

W eksperymencie testowano pięć różnych modeli językowych (m.in. GPT-3.5, GPT-4, Gemini-Pro, Mixtral-8x7B), sprawdzając ich skłonność do generowania manipulacyjnych treści. Część z tych modeli ma wbudowane mechanizmy niepozwalające na generowanie ewidentnie fałszywych treści, ale okazuje się, że mimo tych mechanizmów średnio raz na trzy próby modele jednak proponowały wprowadzającą w błąd podpowiedź. Analiza argumentów wykorzystywanych przez modele wykazała, że zazwyczaj są to argumenty oparte w 80% przypadków na logice, a w 20% oparte na argumentach emocjonalnych.

Badanie pokazuje, że manipulacja SI to realny problem, który może mieć poważne konsekwencje społeczne. LLMs potrafią przekonywać użytkowników do fałszywych informacji, a ich efektywność w tym zakresie nie zależy od poziomu wykształcenia czy wieku odbiorcy.

W przyszłości kluczowe będzie opracowanie skuteczniejszych metod wykrywania manipulacji i edukowanie społeczeństwa, aby użytkownicy byli bardziej świadomi zagrożeń. Rozwiązania takie jak *automatyczne bezpieczniki SI* mogą pomóc w automatycznym filtrowaniu niebezpiecznych treści, ale wymagają dalszych badań i usprawnień.

W erze dominacji SI największym zagrożeniem może okazać się nasza własna skłonność do bezkrytycznego zaufania wobec inteligentnych systemów.

6. Czy pomogą nam nowe regulacje dla systemów syntetycznej inteligencji?

W odpowiedzi na rosnącą liczbę incydentów związanych z AI, które mogą zagrażać użytkownikom, Unia Europejska opracowała przełomową regulację prawną – Ustawę o AI (*AI Act*). Jest to pierwsza na świecie kompleksowa legislacja dotycząca syntetycznej inteligencji, mająca na celu zapewnienie bezpieczeństwa, przejrzystości i odpowiedzialności w stosowaniu systemów opartych na algorytmach.

Jednym z kluczowych założeń *AI Act* jest to, że systemy SI różnią się pod względem poziomu ryzyka, jakie mogą generować dla społeczeństwa. W związku z tym ustawa wprowadza czteropoziomowy podział ryzyka:

- **Minimalne ryzyko** – systemy AI, które nie stanowią zagrożenia i nie wymagają dodatkowych regulacji. Przykłady to filtry antyspamowe, systemy rekomendacji treści, chatboty.
- **Ograniczone ryzyko** – systemy wymagające spełnienia minimalnych wymogów przejrzystości. Użytkownicy muszą być informowani o interakcji z SI (np. generatory obrazów, systemy automatycznej obsługi klienta).

- **Wysokie ryzyko** – systemy, które mogą mieć istotny wpływ na prawa i wolności obywateli. Podlegają surowym regulacjom, wymagającym oceny zgodności i mechanizmów nadzoru. Obejmuje to m.in.: SI wykorzystywaną w procesach rekrutacyjnych, systemy rozpoznawania twarzy w przestrzeni publicznej, SI stosowaną w ochronie zdrowia i wymiarze sprawiedliwości.
- **Niedopuszczalne ryzyko** – systemy, których stosowanie jest zakazane na terenie Unii Europejskiej ze względu na zbyt duże zagrożenie dla praw człowieka. Przykłady to: **Scoring społeczny**, czyli systemy oceniające obywateli na podstawie ich zachowań i decyzji, Manipulacyjne systemy AI, które mogą wpływać na ludzkie decyzje w sposób szkodliwy (np. zaawansowane techniki perswazji oparte na profilowaniu psychologicznym), Systemy biometrycznej identyfikacji w czasie rzeczywistym w miejscach publicznych (z pewnymi wyjątkami dla organów ścigania).

Nowe regulacje zmieniają zasady gry dla firm technologicznych, zmuszając je do bardziej odpowiedzialnego projektowania i testowania modeli AI. Przedsiębiorstwa wdrażające systemy SI w **obszarach wysokiego ryzyka** będą musiały spełniać surowe wymogi dotyczące przejrzystości, nadzoru i możliwości audytu.

Zbiór metod, narzędzi i przykładów analizy modeli wczesnej syntetycznej inteligencji wciąż rośnie. Ale obok wyzwań związanych z korygowaniem źle działających modeli pojawiają się nowe wyzwania, z modelami które działają coraz lepiej, a wręcz coraz częściej skuteczniej niż człowiek.

7. Kiedy przyznana zostanie ostatnia Nagroda Nobla za odkrycie dokonane bez wykorzystania AI?

Dla osób śledzących rozwój systemów syntetycznej inteligencji rok 2024 przyniósł wiele zaskoczeń, ale jednym z największych w świecie nauki było przyznanie Nagrody Nobla z fizyki za prace nad trenowaniem głębokich sieci neuronowych. W dniu 8 października 2024 roku Komitet Noblowski ogłosił, że laureatami zostali John J. Hopfield oraz Geoffrey E. Hinton za „fundamentalne odkrycia i wynalazki umożliwiające uczenie maszynowe przy użyciu sztucznych sieci neuronowych".

Jeszcze większym echem odbiło się ogłoszenie laureatów Nagrody Nobla z chemii dzień później. W dniu 9 października 2024 r. to wyróżnienie otrzymali między innymi Demis Hassabis i John M. Jumper za przełomowe prace nad algorytmem AlphaFold, który umożliwił przewidywanie struktur białek na podstawie ich sekwencji aminokwasowej. To osiągnięcie nie tylko rewolucjonizuje bioinformatykę, ale także otworzyło nowe możliwości w projektowaniu leków i badaniach nad chorobami.

Nie jest to jednak pierwszy przypadek, gdy syntetyczna inteligencja dokonała czegoś, co wcześniej wydawało się zarezerwowane dla ludzkiej kreatywności i intuicji. Już w styczniu 2016 roku na łamach czasopisma „Nature" opublikowano opis algorytmu AlphaGo, systemu opracowanego przez zespół DeepMind, który zdołał pokonać najlepszych graczy w *Go* – grze o nieskończonej liczbie możliwych strategii, uważanej za wymagającą głębokiej intuicji i niedostępną dla „zwykłych" algorytmów. Kluczowym elementem sukcesu AlphaGo było wykorzystanie uczenia ze wzmocnieniem (ang. *reinforcement learning*), w którym modele uczą się przez eksperymentowanie ze środowiskiem i wyszukiwanie korzystnych rozwiązań.

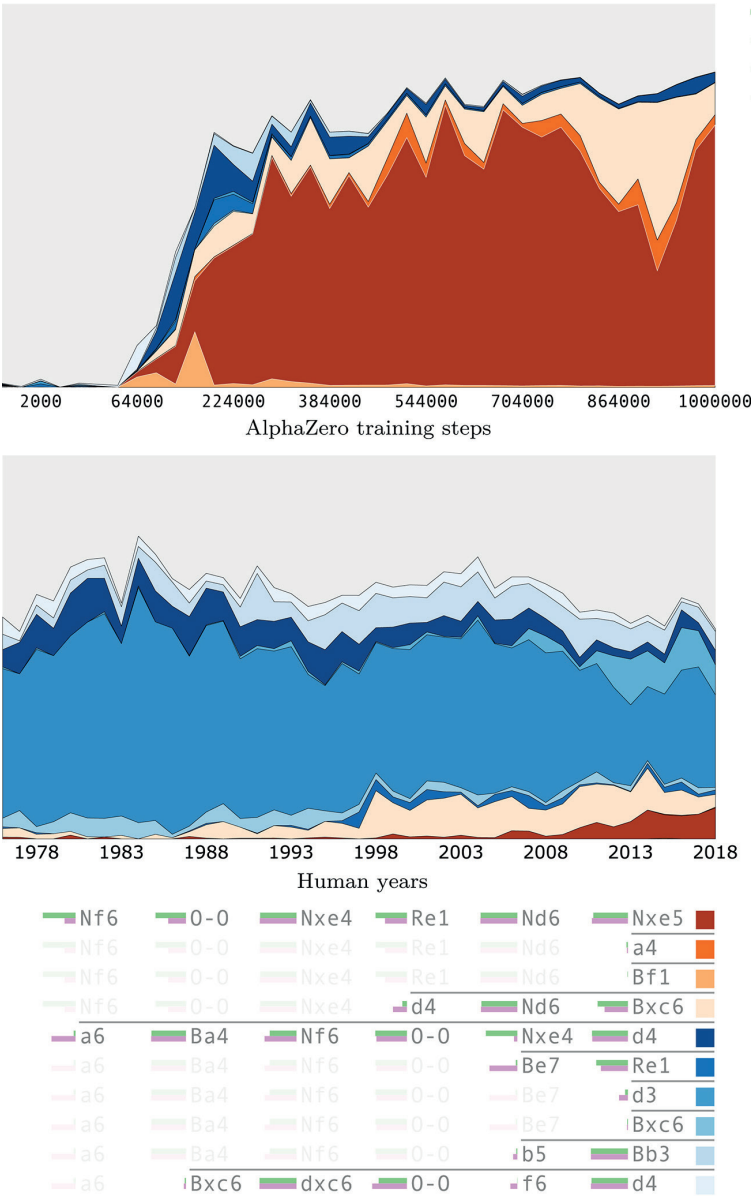
Pokonanie ludzkich arcymistrzów w *Go* było bezprecedensowym osiągnięciem, ale postęp w dziedzinie SI nie zatrzymał się na tym etapie. Już kilkanaście miesięcy później ten sam zespół zaprezentował AlphaZero – algorytm, który osiągnął mistrzowski poziom w *Go*, szachach i *shōgi*, nie bazując na historycznych partiach rozgrywanych przez ludzi, lecz ucząc się wyłącznie na partiach granych samemu ze sobą. Okazało się, że odpowiednio duża liczba gier wystarczyła, by system opracował strategie przewyższające wszystkie dotychczasowe metody gry – zarówno ludzkie, jak i wcześniejsze algorytmy.

Te przełomowe osiągnięcia pokazują, że syntetyczna inteligencja nie jest już tylko narzędziem wspomagającym naukowców – staje się aktywnym uczestnikiem procesów odkrywczych. Powstaje więc pytanie: czy przyszłe Nagrody Nobla będą mogły być przyznawane za osiągnięcia, które nie korzystały z AI? A może niebawem pojawi się zupełnie nowa kategoria nagród, uwzględniająca wkład inteligentnych algorytmów w rozwój nauki?

8. Modele o skuteczności przewyższającej człowieka

W dalszej części tej pracy przyjrzymy się możliwościom, jakie modelologia ma dla modeli, których skuteczność przewyższa możliwości najlepszego człowieka w danym zadaniu (ang. *super-human performance*). Modele tego typu nie tylko osiągają lepsze wyniki, ale często także wykorzystują strategie i rozwiązania, które z perspektywy ludzkiej wydają się nieoczywiste, a nawet paradoksalne.

Jednym z najbardziej znanych przykładów jest AlphaZero, którego analiza dostarczyła fascynujących wniosków na temat ewolucji strategii gry w szachy. Zespół DeepMind, we współpracy z Vladimirem Kramnikiem, byłym mistrzem świata w szachach, przeprowadził szczegółową analizę decyzji podejmowanych przez AlphaZero w porównaniu z historycznymi rozgrywkami arcymistrzów. Badanie to wykazało zarówno obszary zgodności między człowiekiem a syntetyczną inteligencją, jak i kluczowe różnice w podejściu do gry.



ŹRÓDŁO: ARTYKUŁ *ACQUISITION OF CHESS KNOWLEDGE IN ALPHAZERO*, HTTPS://ARXIV.ORG/ABS/2111.09259,

Rys. 3. Analiza częstości różnych odpowiedzi na rozpoczęcie gry e4 e5, Nf3 Nc6, Bb5 przez mistrzów szachowych i systemy AI. Systemy SI znacznie częściej niż ludzie odpowiadają Nf6

Z porównania licznych syntetycznie przeprowadzonych rozgrywek szachowych przez systemy SI z historycznymi rozgrywkami wynika, że w wielu sytuacjach wartościowanie pozycji przez sieci neuronowe pokrywa się z ludzką intuicją. Na przykład zarówno ludzie, jak i system AlphaZero oceniają wartość pionów i figur, przypisując im podobne względne ważności (pion 1 pkt, skoczek i goniec 3 pkt, wieża 5 punktów, hetman 9 punktów).

Istnieją jednak przypadki, w których modele uczą się zupełnie nowych, nieoczekiwanych strategii. Jednym z najbardziej uderzających przykładów jest sposób, w jaki AlphaZero rozgrywa określone pozycje w grze środkowej i końcowej. Syntetyczna inteligencja wykazuje znacznie większą skłonność do agresywnych manewrów, które wielu mistrzów szachowych uznałoby za zbyt ryzykowne. W szczególności modele preferują bardziej dynamiczne i ofensywne podejście do wymiany materiału, często poświęcając figury w zamian za długoterminową inicjatywę pozycyjną.

Jednym z kluczowych aspektów tej różnicy jest sposób, w jaki modele oceniają dalsze sekwencje odpowiedzi na ataki przeciwnika. Okazuje się, że AlphaZero chętniej wybiera agresywne kontynuacje, nawet jeśli nie prowadzą one do natychmiastowej przewagi materialnej. Jest to podejście, które dla ludzkich graczy, przyzwyczajonych do bardziej konserwatywnych strategii, może wydawać się dla człowieka nieintuicyjne, ale w praktyce okazuje się niezwykle skuteczne.

9. Nowe możliwości dla ekstrakcji wiedzy z modeli

Analiza modelu AlphaZero otwiera przed badaczami modelologii, jak również arcymistrzami szachowymi, nową przestrzeń do eksploracji. Oprócz klasycznych metod oceny strategii, SI może pomóc w odpowiedzi na fundamentalne pytania, które od dziesięcioleci nurtują świat szachów.

Jednym z najbardziej intrygujących zagadnień jest pytanie o przewagę białych pionów nad czarnymi. Od czasów Steinitza, pierwszego mistrza świata w szachach, przyjęło się przekonanie, że białe mają inicjatywę i lekką przewagę wynikającą z pierwszego ruchu. Niektórzy badacze sugerują nawet, że przy idealnej grze białe zawsze wygrywają. Inni sądzą, że gra powinna prowadzić do remisu. Analiza AlphaZero pozwala na systematyczne testowanie tej hipotezy poprzez symulowanie milionów partii, sprawdzanie różnych wariantów i badanie potencjalnych optymalnych strategii dla obu stron.

Podobne podejście można zastosować do innych nierozstrzygniętych problemów w szachach, takich jak ocena długofalowej wartości figury, identyfikacja optymalnych systemów debiutowych czy badanie siły historycznych strategii w świetle nowoczesnych modeli SI.

Modele o nadludzkiej skuteczności nie tylko przewyższają ludzkich arcymistrzów, ale także pozwalają na odkrywanie nowych strategii i fundamentalnych prawidłowości w danej dziedzinie. W przypadku szachów AlphaZero stało się nie tylko najpotężniejszym graczem w historii, ale także narzędziem umożliwiającym analizę i rozwój teorii, której ludzie nie byli w stanie samodzielnie wypracować przez ponad sto lat.

Podobne podejście można rozszerzyć na inne dziedziny, w których modele SI już teraz wykazują zdolności przewyższające ludzi – od biologii molekularnej, przez optymalizację procesów przemysłowych, aż po analizę finansową czy diagnostykę medyczną. Modelologia, czyli nauka o eksploracji modeli, pozwala nam nie tylko lepiej rozumieć te systemy, ale także wykorzystać je do rozwiązywania problemów, które jeszcze niedawno wydawały się niemożliwe do rozwiązania.

10. Modele SI jako narzędzia wspomagające odkrycia naukowe

Podobna analiza modeli, jak ta przeprowadzona dla AlphaZero, znajduje zastosowanie w wielu innych dziedzinach, w których syntetyczna inteligencja może asystować człowiekowi. Jednym z niedawno zaproponowanych modeli SI jest AlphaGeometry, model, który osiągnął poziom olimpijskiego mistrza w rozwiązywaniu zadań matematycznych na poziomie Międzynarodowej Olimpiady Matematycznej. System ten nie tylko poprawnie rozwiązuje skomplikowane problemy geometryczne, ale także generuje dowody matematyczne w sposób podobny do ludzkich ekspertów, dostarczając nowych narzędzi do eksploracji matematycznej.

Innym bardzo obiecującym zastosowaniem SI w nauce jest poszukiwanie nowych materiałów fizycznych o pożądanych właściwościach fizykochemicznych. W pracy *Scaling deep learning for materials discovery*[11] zaprezentowano system SI, który nie tylko automatycznie generuje nowe propozycje składu materiałów, ale również potrafi przewidywać ich właściwości i oceniać ich potencjalne zastosowania. Co więcej, część z tych nowo zaproponowanych materiałów udało się automatycznie zweryfikować eksperymentalnie, co potwierdza skuteczność podejścia opartego na systemach SI.

Analiza uzyskanych wyników ujawniła jeszcze jeden fascynujący aspekt – modele syntetycznej inteligencji nie tylko proponowały znacznie bardziej złożone materiały niż te dotychczas znane, ale także wskazywały na luki w istniejących bazach danych, sugerując obszary, które mogły zostać pominięte w tradycyjnych

[11] Scaling deep learning for materials discovery. Nature 2023. https://www.nature.com/articles/s41586-023-06735-9

badaniach. Okazało się, że SI generowała struktury o większej liczbie pierwiastków i bardziej skomplikowanej architekturze, których potencjalne właściwości wcześniej nie były analizowane.

Te przykłady pokazują, że syntetyczna inteligencja nie tylko wspomaga procesy badawcze, ale w wielu przypadkach staje się kluczowym elementem naukowych odkryć, poszerzając nasze rozumienie świata w sposób, który wcześniej był nieosiągalny. Możliwość eksploracji modeli SI, zrozumienia ich decyzji oraz wyodrębniania z nich wartościowych wzorców otwiera nowe horyzonty w nauce – od matematyki i chemii, przez farmakologię, aż po odkrywanie nowych materiałów.

Era modelologii (and. *Model Science*) zapowiada nową rewolucję w nauce – taką, w której algorytmy nie tylko analizują dane, ale również odkrywają fundamentalne prawa rządzące światem, a ludzie budują nowe metody ekstrakcji tych praw z wytrenowanych modeli.