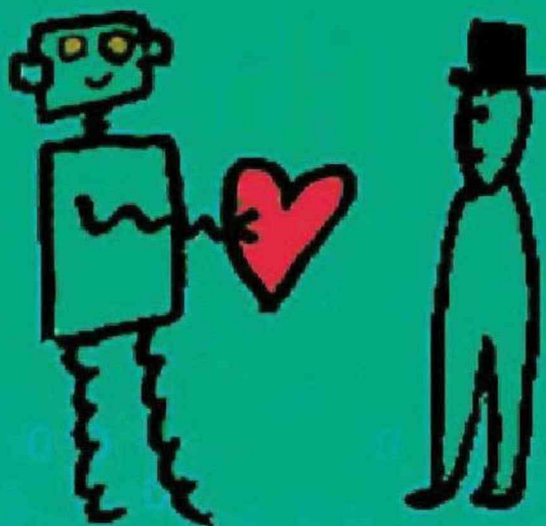


NAUKA PROJEKTPULSAR.PL

ILUSTRACJA PATRYK SROCZYŃSKI



Ucieczka z ewolucji

Jak symulować emocje w maszynie? Jak wzbudzać zaufanie do bytów cyfrowych?
I w jakim celu? Rozmowa z Janem K. Argasińskim, medioznawcą, filozofem
i neuroinformatykiem z **Uniwersytetu Jagiellońskiego**.

KATARZYNA CZARNECKA: – **Jak to jest w grach komputerowych: pies merda ogonem czy ogon macha psem?**

JAN K. ARGASIŃSKI: – Pies merda ogonem. **Mówi pan jako programista i twórca?** Tak.

A jak odpowiedziałby pan jako użytkownik?

Że mogę w grze robić, co chcę, na tyle, na ile pozwala mi jej świat. A to jest definiowane przez twórców. Ucząc studentów projektowania, mówię o „iluzji swobody”. Użytkownik ma mieć poczucie kontroli. A to nie jest to samo, co jej rzeczywiste posiadanie. Bo im większą swobodę twórcy dadzą graczowi, tym trudniej im

kontrolować fabułę. Zawężają więc pulę wyborów, których można w grze dokonać, choć może nie aż tak bardzo, jak pisarz ogranicza czytelnikowi możliwości interakcji z treścią książki. Istnieją oczywiście gry, w których celem jest jedynie manipulacja światem wokół gracza, a nie opowiadanie historii. I to też może być satysfakcjonujące, choć ja takie gry nazywam „zabawkami”. Bo są jak klocki Lego – układamy je wedle własnej fantazji. Gra ma strukturę – wokół fabuły, rywalizacji czy co tam sobie twórcy wymyślili. Z tego punktu widzenia system jest psem, a gracz – ogonem. Poza tym gra może istnieć bez gracza, ale on bez niej nie.

Człowiek musi się podporządkować twórcowi cyfrowemu.

Może też zrezygnować z gry. Jeśli się jednak na nią decyduje, to musi wejść w jej realia. Jak z książką: czytając, zgadzamy się na świat przedstawiony, a jeżeli nam nie odpowiada, rezygnujemy z lektury. W pewnym sensie gra może być literaturą rozszerzoną.

Komunikacja w niej może być obustronna?

Na pewno jest interaktywna. To, co robi gracz, ma znaczenie dla świata przedstawionego. Żeby była obustronna, po drugiej stronie musi być podmiot, czyli inny ludzki gracz. W przypadku ►

NAUKA PROJEKTPULSAR.PL

► *singleplayer* go nie ma. To sytuacja człowieka kontra gra.

Na pewno? A nie zależy to od tego, w jaki sposób człowiek postrzega grę?

Można tak spojrzeć. W każdej, nawet międzyludzkiej komunikacji projektujemy sobie osobę po drugiej stronie. Rozmawiając z komputerem, telefonem czy lodówką, w jakimś stopniu je uczłowiczamy. A kto się nigdy nie wkurzył na laptopa, bo uznał jakiś komunikat za złośliwy, niech pierwszy rzuci kamieniem.

To wkurzenie (*computerrage*) często kończy się dla sprzętów źle.

W grach jest to *rage quit*, kiedy ktoś rzuca kontrolerem i wychodzi. Zdarza się tak, bo jest to komunikacja traktowana bardzo serio.

Pan chciałby ją zmodernizować.

Chodzi o *affective computing*, przetwarzanie emocjonalne. Urządzenie ma odczytywać emocje gracza. Możliwość wglądu w uczucia innego człowieka dają słowa, ale i sygnały cielesne. Jednak nasze emocje są dostępne tylko dla nas. U innych obserwujemy zewnętrzne ich wyrazy. To komunikacja przez empatię, którą każdy z nas praktykuje na co dzień. Mój zespół próbował nauczyć systemy elektroniczne odczytywać sygnały fizjologiczne, takie jak tętno czy aktywność elektryczna mózgu, i interpretować je w świetle afektów.

Z sukcesem?

Połowicznym. Jesteśmy w stanie to robić, ale urządzenia rejestrujące sprawiają problemy. Na przykład elektroencefalografia (EEG) jest podatna na szumy. Jeśli nie ma ekranowania, jak w laboratorium, a jeszcze badany podskakuje czy robi przysiady, jak to bywa w grach w wirtualnej rzeczywistości, to sygnał może być katastroficznie złej jakości. Drugi kłopot polega na tym, że jest bardzo wiele subtelnych sposobów, w jaki ciało okazuje uczucia. Kiedy się w dodatku na siebie nakładają, są trudne do zidentyfikowania. No i my sami mamy problem z ich klasyfikacją, bo często są do siebie podobne.

Lisa Feldman Barrett z Northeastern University, Harvard Medical School i Massachusetts General Hospital, która prowadzi badania w dziedzinie psychologii i neuronauki, w książce „Jak powstają emocje” opisała swój przypadek. Miała na studiach kolegę, który wyrażał wobec niej zainteresowanie romantyczne. Bez wzajemności. Któregoś dnia dała mu się jednak zaprosić na spotkanie. W pewnym momencie poczuła szybsze bicie serca, zrobiło jej się

ciepło itd. Pomyślała więc, że może jednak coś do niego czuje. A później wróciła do domu i resztę wieczoru spędziła wymiotując, ponieważ miała grypę. Nawet osoba studiująca psychologię dość łatwo może pomylić chorobę z afektem.

I że rozumienie sygnałów zależy od kontekstu.

Zgadza się. My budowaliśmy systemy, które teoretycznie go znały. Wiedzieliśmy, że w danym momencie gry użytkownika straszmy, więc zakładaliśmy, że kiedy w odpowiednich parametrach fizjologicznych widzimy pobudzenie, to ono jest związane ze strachem. I mogliśmy za pomocą naszego inteligentnego systemu modyfikować przebieg rozgrywki. A nawet dostarczać odpowiednich bodźców, by pokierować emocjami odbiorcy. Nie byliśmy jednak całkowicie zadowoleni z efektów.

Zastosowanie jednak jest.

Tak, w symulacjach szkoleniowych dla ratowników czy strażaków. Temat jest niesłychanie złożony i minie jeszcze sporo czasu, zanim to będzie naprawdę dobrze robione. Dodatkowym celem naszego projektu było zdiagnozowanie takiego parametru jak zaangażowanie w rozgrywkę. I tu już przede wszystkim chodziło o możliwość manipulacji emocjami użytkownika.

Nie brzmi to dobrze.

Jest jednak nieodłączną cechą gier, szczególnie rozrywkowych. W nich przecież chodzi o to, żeby kogoś przestraszyć, zachwycić itd., a walutą – i to bardzo cenną – jest właśnie zaangażowanie: czas spędzony w rozgrywce i pieniądze wydane na nią w realu. Poza tym otaczają nas dziś urządzenia, które potencjalnie zbierają informacje na temat stanów emocjonalnych. Smartwatch wie, czy jesteśmy w kinie, czy w szpitalu. Po lokalizacji albo dlatego, że mamy to wpisane w kalendarzu. A wyszukiwarki internetowe? Z danych, których codziennie dostarczamy, odczytują kontekst. Algorytmy wykorzystują wiedzę o nas. Systemów, które kierują się sygnałami fizjologicznymi i czytają nasze uczucia, nie ma wiele. Jeszcze nie ma.

Nie ma też wielu takich, które realizują drugą część *affective computing*.

Zapewne także „jeszcze”?

Chodzi o to, że system ma symulować emocje. Człowiek chce komunikować się z kimś, kto jest jakiś. Bez tego nie da się zbudować więzi czy zaufania. Jeżeli człowiek się na nas krzywi albo się do nas uśmiecha, to wywołuje u nas reakcje. Tak samo się dzieje, kiedy wiemy, że byt po drugiej

stronie jest sztuczny. Chcemy, żeby potrafił nas czytać i żebyśmy widzieli jakieś jego domniemane stany wewnętrzne.

Trzeba dać mu osobowość?

Oczywiście. Żeby nam było przyjemnie.

I tego właśnie też pan

ze współpracownikami próbował.

W kwietniu ukazała się wasza książka.

Stworzyliśmy z Karolem Przystalskim, Natalią Lipp i Dawidem Pacholczykiem „neurotycznego ChataGPT”. A także intrygującego i narcystycznego. I inne modele językowe też poddawaliśmy takim modyfikacjom.

A zatem na scenę wkracza sztuczna inteligencja.

Owszem. W przypadku dużych modeli językowych (LLM) medium jest – jak nazwa wskazuje – język, najczęściej pisany. I nawet jeśli zmienimy go na mowę, to nadal jest to komunikacja, z której zostają wycięte wszystkie kwestie fizjologiczne, mimiczne itd. Tekst jednak też może być nacechowany emocjonalnie. LLM są zaś dialogiczne: my o coś prosimy, one nam to dają. Możemy zatem – i to właśnie zrobiliśmy – modyfikować je tak, żeby odpowiedzi były zgodne z pewnym profilem osobowości, z którym chcielibyśmy mieć do czynienia. Mało tego – nasz LLM generuje też innego rodzaju treści, choćby wizualne. Na przykład portrety Chopina stworzone tak, jakby ich autorzy mieli różne typy osobowości.

Powiedział pan, że takie zabiegi zwiększają zaufanie. Tyle że przecież LLM halucynują, czyli po prostu kłamią. Wiadomo, że nie można im wierzyć.

To zależy.

Ulubiona kwestia naukowca.

Przyznaję. Chcę tylko powiedzieć, że to, czy należy im ufać, to jest osobny temat. My w Sano – Centrum Zindywidualizowanej Medycyny Obliczeniowej, działamy na AI wspomagającej różnego rodzaju decyzje medyczne. Przede wszystkim zależy nam, żeby nasz wirtualny agent był odpowiednio przygotowany merytorycznie. Nawet jeśli niekoniecznie będzie diagnozował, ale tylko udzielał wskazówek dotyczących trybu życia pacjenta. Chcemy też, żeby lekarz lub pacjent mu ufał. Próbuje więc budować to zaufanie. Oczywiście możemy też osiągnąć sytuację, kiedy taki mechanizm będzie przez ludzi używany w złych celach. Jak teraz wygenerowane przez AI zdjęcia głodnych dzieci, które rzekomo potrzebują wsparcia finansowego.

Powinniśmy zatem móc jakoś odróżnić Donalda Trumpa od prawdziwego papieża.

Ludzkość biologicznie nie ewoluuje tak szybko jak technologie. My nie zmieniliśmy się w ciągu ostatniego tysiąca lat za bardzo. Na tworzone za pomocą mediów iluzje jesteśmy bardzo podatni. Nawet jeśli mamy świadomość ich źródła. Przykładem są gry czy kino. To tylko płaskie obrazy wyświetlane „na prześcieradle”, a my inwestujemy w nie tony emocji.

No tak, ale czy można stworzyć jakieś narzędzia, które umożliwią odróżnienie, co jest prawdziwe, a co nie?

One się powolutku pojawiają, choć to bardzo trudne. Teraz kończymy artykuł naukowy na temat możliwości detekcji źródeł komunikatów tekstowych (wykrywania LLM). Alternatywą są rozwiązania hybrydowe, certyfikacja i tzw. wyjaśnialność modeli (*explainability*). Bo kiedy ktoś mi mówi: „Masz tu świetną aplikację, to jest wirtualny lekarz, powiedz mu, jak się czujesz, on ci postawi diagnozę”, nie pytam, czy to tylko dobry system informatyczny. Interesuje mnie, kto go stworzył, na jakich danych się uczył i kto się pod nim podpisał. Możliwość zidentyfikowania eksperckiego autorytetu może być jedną z naszych kotwic w rzeczywistości.

Eksperci się niepokoją, że niedługo ludzie będą żyć w systemie, którego nie rozumieją i na który nie mają żadnego wpływu. Będą w nim tylko wkładką mięsną realizującą wytyczne AI. To przesada?

Jest taka obawa. Z tym że ta wkładka mięsna wydaje się niezbędna. LLM są trenowane na tekstach wyprodukowanych przez ludzi, a te się powoli kończą, choćby dlatego, że mamy w internecie coraz więcej nieoznaczonych odpowiedzi treści generowanych przez AI. Jeśli zaś zacznie się trenować nowe modele językowe na treściach stworzonych przez stare, to cała ta konstrukcja się posypie. Na marginesie: kiedy robimy badania stylometryczne, czyli porównujemy teksty wygenerowane z innymi, to musimy korzystać z artykułów naukowych sprzed np. 2005 r., bo późniejsze są niepewne. Wyobrażam sobie dziś taki matrix, w którym ludzie będą siedzieć i pisać, tworzyć – cokolwiek, byle było wytworem biologicznego układu nerwowego – żeby mieć czym nakarmić AI.



Dr Jan K. Argasiński pracuje na Wydziale Fizyki, Astronomii i Informatyki Stosowanej Uniwersytetu Jagiellońskiego. Zajmuje się neurobiologią obliczeniową i biologicznie inspirowaną sztuczną inteligencją, a także problematyką przetwarzania emocji w systemach informatycznych oraz wirtualną i poszerzoną rzeczywistością. W Sano – Centrum Medycyny Obliczeniowej prowadzi grupę Computational Neuroscience.

A teraz poważnie: kluczowe jest to, że mamy ręce, nogi, dotyk, wzrok, słuch, smak i węch oraz rzeczywistość, która nam dostarcza nowych bodźców. Na nich nadbudujemy nowe połączenia neuronowe, czyli uczymy się nowych rzeczy i nowych emocji. Właśnie to, że świat jest dla nas ciągle zaskakujący, pozwala nam tworzyć nowe treści, podczas gdy takie LLM to tylko słoiki, do których coś tam wrzucamy i kisimy. Swoją drogą kolejnym krokiem w emancypacji sztucznej inteligencji mogłoby być wyposażenie jej w ciało i wypuszczenie w świat. Niech się w nim uczy, niech go odczuwa. Wtedy może będzie miała prawdziwe emocje. Musimy ostatecznie zrobić coś na nasz obraz i podobieństwo.

Już zrobiliśmy. Nawet w Polsce

– robot Clone Alpha ma sztuczne mięśnie, ścięgna i kości oraz „układ krwionośny”. Wymyślono go do prac domowych, ale ponoć jest także dowcipny. Przed sprzedaż ma ruszyć w tym roku. Jakiś matrix już się zbliża?

Raymond Kurzweil wydał pod koniec XX w. książkę „Nadchodzi osobliwość”, a rok temu ukazała się „Osobliwość coraz bliżej”, w której twierdzi, że wszystko, co napisał, na razie się sprawdza.

„Człowiek zintegruje się z technologią, dając początek nowej formie życia gatunkowego” – tak przedstawiał jego wniosek Edwin Bendyk w POLITYCE 33/24.

Tak. Jest jednak alternatywa: te sztuczno-inteligentne byty będą do nas podobne, ale pójdą w swoją stronę. A gdyby nawet powstała w nich niezależna świadomość, to skąd mielibyśmy to wiedzieć? I dlaczego mielibyśmy akurat tego się bać?

Ze względu na podobieństwo fizyczne.

Teraz pod tym względem nie radzimy sobie nawet ze zwierzętami. Zakładamy, że sympanz czy orangutan jako nam pokrewny ma jakąś świadomość. A co z ośmiornicą, gatunkiem na pewno inteligentnym, ale radykalnie innym od nas? Czy dopuszczamy do siebie myśl, że istnieje świadomość ośmiornicza? Po samej złożoności układu nerwowego moglibyśmy zakładać, że tak, ale jak się

do tego ustosunkować? Trzeba byłoby się zastanowić, czy to nie jest dla nas zobowiązanie etyczne.

Kurt Vonnegut stwierdził kiedyś:

„Uczono mnie, że ludzki umysł jest szczytowym osiągnięciem ewolucji, ale wydaje mi się, że to kiepski patent na przetrwanie”. A może jednak nie miał racji, tylko droga będzie inna?

Ewolucja służy wyłącznie do tego, żeby przekazać geny dalej. Są jednak myśliciele, którzy twierdzą, że my, ludzie, dzięki technologiom nawet tak podstawowym jak pismo, wyjęliśmy się nieco z ewolucyjnego paradygmatu. Dysponujemy technologią, która pozwala nam dowolną niszę zająć i się do niej zaadaptować. Zamiast ewolucji genetycznej mamy ewolucję idei, zamiast genów mamy memy. I AI może być takim memem, taką ideą, która pozwoli nam trwać. Wyrwać się zupełnie z prostej biologicznej ewolucji. To głosi posthumanizm.

I co pan na to?

Powiedziałbym: „Nie tak szybko”. Bo jednak chorujemy, umieramy, więc nie jesteśmy aż tak bardzo poza biologią. Czy jednak AI nie może być przedłużeniem gatunku ludzkiego? Bo co to jest gatunek ludzki? Czy to są te ciała fizyczne, ten byt mięsny co do zasady stojący w pionie? A może rzeczywiście są to idee? Możemy sobie wyobrazić taką wizję świata, w której ludzie wyginęli, a one zostały w AI.

To może na wszelki wypadek ona już powinna mieć prawa człowieka?

Krzywdząc coś, co jest do nas podobne, w sensie etycznym robimy też krzywdę sobie. Czy to jednak jest to złe dla AI, jakiegoś kawałka oprogramowania, które sobie funkcjonuje na serwerze? Wydaje mi się, że jeszcze za wcześnie, żeby się o to martwić. Wciąż jeszcze AI jest ogonem, a człowiek psem.

ROZMAWIAŁA
KATARZYNA CZARNECKA

Więcej na ten temat podczas tegorocznego Copernicus Festival (20–25 maja w Krakowie), którego patronami medialnymi są POLITYKA i serwis pulsar: copernicusfestival.com