

Deklaracja Bletchley przyjęta przez kraje uczestniczące w Szczycie Bezpieczeństwa AI, 1-2 listopada 2023 roku

1 listopada 2023 roku

Sztuczna Inteligencja (SI) stwarza ogromne globalne możliwości: ma potencjał, by przekształcić i poprawić dobrobyt, pokój oraz pomyślność ludzkości. Aby to osiągnąć, potwierdzamy, że w interesie ogółu SI powinna być projektowana, rozwijana, wdrażana i wykorzystywana w sposób bezpieczny, skoncentrowany na człowieku, godny zaufania i odpowiedzialny. Z zadowoleniem przyjmujemy dotychczasowe wysiłki społeczności międzynarodowej na rzecz współpracy w dziedzinie SI, mające na celu promowanie inkluzywnego wzrostu gospodarczego, zrównoważonego rozwoju i innowacji, ochronę praw człowieka i podstawowych wolności, a także budowanie zaufania publicznego do systemów SI, aby w pełni zrealizować ich potencjał.

Systemy SI są już stosowane w wielu dziedzinach codziennego życia, takich jak mieszkalnictwo, zatrudnienie, transport, edukacja, zdrowie, dostępność oraz wymiar sprawiedliwości, a ich zastosowanie prawdopodobnie będzie się rozszerzać. Uznajemy, że to wyjątkowy moment na podjęcie działań oraz na potwierdzenie potrzeby bezpiecznego rozwoju SI i wykorzystania transformacyjnych możliwości SI w sposób inkluzywny, służący dobru ogółu w naszych krajach oraz na świecie. Dotyczy to także usług publicznych, takich jak zdrowie i edukacja, bezpieczeństwo żywnościowe, nauka, czysta energia, bioróżnorodność i klimat, by umożliwić korzystanie z praw człowieka oraz wzmocnić wysiłki na rzecz realizacji Celów Zrównoważonego Rozwoju ONZ.

Wraz z tymi możliwościami SI niesie również ze sobą istotne zagrożenia, w tym w obszarach codziennego życia. W związku z tym z zadowoleniem przyjmujemy międzynarodowe wysiłki mające na celu badanie i przeciwdziałanie potencjalnym skutkom systemów SI w istniejących forach oraz innych odpowiednich inicjatywach. Doceniamy znaczenie ochrony praw człowieka, przejrzystości i wyjaśnialności, uczciwości, odpowiedzialności, regulacji, bezpieczeństwa, właściwego nadzoru ludzkiego, etyki, ograniczania uprzedzeń, ochrony prywatności i danych osobowych. Zauważamy również potencjalne ryzyko wynikające z możliwości manipulowania treściami lub generowania wprowadzających w błąd informacji. Wszystkie te kwestie są niezwykle ważne, dlatego podkreślamy konieczność i pilność ich rozwiązania.

Szczególne ryzyko związane jest z tzw. „graniczną” SI, rozumianą jako wysoce zaawansowane modele ogólnego przeznaczenia, w tym modele bazowe, które mogą wykonywać szeroki zakres zadań, a także z wąsko wyspecjalizowaną SI, która może wykazywać zdolności prowadzące do szkód. Takie ryzyko może wynikać z celowego

niewłaściwego wykorzystania lub niezamierzonych problemów związanych z dostosowaniem tych systemów do ludzkich intencji. Problemy te są częściowo spowodowane tym, że możliwości tych systemów nie są w pełni zrozumiałe, a ich skutki trudne do przewidzenia. Szczególnie obawiamy się takich zagrożeń w dziedzinach takich jak cyberbezpieczeństwo i biotechnologia, a także tam, gdzie systemy granicznej SI mogą potęgować ryzyko, takie jak dezinformacja. Istnieje potencjał poważnych, a nawet katastrofalnych szkód, zarówno zamierzonych, jak i niezamierzonych, wynikających z najbardziej zaawansowanych zdolności tych modeli SI. Biorąc pod uwagę szybkie i nieprzewidywalne tempo rozwoju SI oraz wzrost inwestycji w technologie, podkreślamy pilną potrzebę pogłębiania naszej wiedzy na temat tych potencjalnych zagrożeń i działań mających na celu ich rozwiązanie.

Wiele zagrożeń związanych z SI ma charakter międzynarodowy, co oznacza, że najlepiej można je rozwiązać poprzez współpracę międzynarodową. Zobowiązujemy się do współdziałania w sposób inkluzywny, aby zapewnić, że SI będzie skoncentrowana na człowieku, godna zaufania, odpowiedzialna i bezpieczna, służąc dobru ogółu. W tym celu zamierzamy korzystać z istniejących międzynarodowych forów oraz innych odpowiednich inicjatyw, by wspierać współpracę w obszarze szerokiego spektrum zagrożeń związanych z SI. W tym kontekście uznajemy, że państwa powinny rozważyć znaczenie proporcjonalnego podejścia do regulacji i zarządzania, które sprzyja innowacjom i maksymalizuje korzyści, uwzględniając jednocześnie ryzyka związane z SI. Może to obejmować, tam gdzie jest to stosowne, wprowadzanie klasyfikacji i kategoryzacji ryzyk w oparciu o krajowe okoliczności oraz obowiązujące ramy prawne. Zwracamy również uwagę na znaczenie współpracy, tam gdzie to właściwe, w zakresie wspólnych zasad i kodeksów postępowania.

W odniesieniu do specyficznych zagrożeń najczęściej związanych z graniczną SI zobowiązujemy się intensyfikować i utrzymywać naszą współpracę oraz rozszerzać ją na kolejne kraje w celu identyfikacji, zrozumienia i, tam gdzie to odpowiednie, podejmowania działań poprzez istniejące międzynarodowe fora i inne odpowiednie inicjatywy, w tym przyszłe międzynarodowe Szczyty Bezpieczeństwa SI.

Wszyscy uczestnicy mają rolę do odegrania w zapewnieniu bezpieczeństwa SI: państwa, fora międzynarodowe, inicjatywy, przedsiębiorstwa, społeczeństwo obywatelskie i środowisko akademickie muszą współpracować. Podkreślając znaczenie inkluzywnego podejścia do SI i zmniejszania przepaści cyfrowej, ponownie potwierdzamy, że współpraca międzynarodowa powinna dążyć do zaangażowania szerokiego grona partnerów, tam gdzie to właściwe. Z zadowoleniem przyjmujemy polityki i podejścia zorientowane na rozwój, które mogą pomóc krajom rozwijającym się w budowie zdolności w zakresie SI oraz w wykorzystaniu jej potencjału na rzecz zrównoważonego wzrostu i zmniejszania luki rozwojowej.

Potwierdzamy, że choć bezpieczeństwo należy uwzględniać na wszystkich etapach cyklu życia SI, to podmioty rozwijające graniczne zdolności SI, w szczególności systemy SI, które są wyjątkowo potężne i potencjalnie szkodliwe, mają szczególnie silną odpowiedzialność za zapewnienie ich bezpieczeństwa. Obejmuje to m.in. testy bezpieczeństwa, oceny oraz inne odpowiednie środki. Zachęcamy wszystkie właściwe podmioty do zapewnienia odpowiedniego poziomu przejrzystości i odpowiedzialności w kontekście ich planów dotyczących pomiaru, monitorowania i łagodzenia potencjalnie szkodliwych zdolności oraz związanych z nimi skutków, które mogą się pojawić, zwłaszcza w celu zapobiegania niewłaściwemu wykorzystaniu, problemom z kontrolą oraz eskalacji innych zagrożeń.

W kontekście naszej współpracy oraz w celu informowania o działaniach na poziomie krajowym i międzynarodowym nasza agenda dotycząca ryzyk związanych z graniczną SI będzie koncentrować się na:

- Identyfikacji ryzyk związanych z bezpieczeństwem SI, które budzą wspólne obawy, budowaniu wspólnego, naukowego i opartego na dowodach zrozumienia tych ryzyk oraz podtrzymywaniu tego zrozumienia w miarę rozwoju zdolności SI, w szerszym kontekście globalnych wysiłków na rzecz zrozumienia wpływu SI na nasze społeczeństwa.
- Tworzeniu odpowiednich polityk opartych na ocenie ryzyka w naszych krajach, aby zapewnić bezpieczeństwo w obliczu takich zagrożeń, współpracując tam, gdzie to właściwe, z uznaniem, że nasze podejścia mogą się różnić w zależności od krajowych okoliczności i obowiązujących ram prawnych. Obejmuje to, oprócz zwiększenia przejrzystości ze strony podmiotów prywatnych rozwijających graniczne zdolności SI, opracowanie odpowiednich wskaźników oceny, narzędzi do testowania bezpieczeństwa, a także rozwój odpowiednich zdolności sektora publicznego oraz badań naukowych.

W celu realizacji tej agendy zobowiązujemy się wspierać międzynarodową, inkluzywną sieć badań naukowych nad bezpieczeństwem granicznej SI, która obejmuje i uzupełnia istniejącą oraz nową współpracę wielostronną, plurilateralną i bilateralną, w tym poprzez istniejące międzynarodowe fora i inne odpowiednie inicjatywy. Ma to na celu dostarczanie najlepszej dostępnej wiedzy naukowej, która może być wykorzystywana w procesach politycznych i dla dobra publicznego.

Uznając transformacyjny pozytywny potencjał SI oraz w ramach zapewniania szerszej międzynarodowej współpracy w zakresie SI, zobowiązujemy się do utrzymania inkluzywnego globalnego dialogu, który angażuje istniejące międzynarodowe fora i inne odpowiednie inicjatywy oraz w otwarty sposób przyczynia się do szerszych międzynarodowych dyskusji. Zobowiązujemy się również do kontynuowania badań nad bezpieczeństwem granicznej SI, aby zapewnić, że korzyści płynące z tej technologii będą wykorzystywane odpowiedzialnie, dla dobra i na rzecz wszystkich.

Z niecierpliwością oczekujemy na kolejne spotkanie w 2024 roku.

Porozumienie

Deklarację podpisały następujące kraje:

- Australia
- Brazylia
- Kanada
- Chile
- Chiny
- Unia Europejska
- Francja
- Niemcy
- Indie
- Indonezja
- Irlandia
- Izrael
- Włochy
- Japonia
- Kenia
- Królestwo Arabii Saudyjskiej
- Niderlandy
- Nigeria
- Filipiny
- Republika Korei
- Rwanda
- Singapur
- Hiszpania
- Szwajcaria
- Turcja
- Ukraina
- Zjednoczone Emiraty Arabskie
- Zjednoczone Królestwo Wielkiej Brytanii i Irlandii Północnej
- Stany Zjednoczone Ameryki

Odniesienia do „rządów” i „państw” obejmują organizacje międzynarodowe działające zgodnie z ich kompetencjami legislacyjnymi lub wykonawczymi.