Czy ogólna sztuczna inteligencja przyniesie koniec ludzkości czy „róg obfitości”? Modelowanie gospodarki w warunkach technologicznej osobliwości

Ostatnimi laty obserwujemy ogromne przyspieszenie prac nad algorytmami sztucznej inteligencji. Sztuczna inteligencja (AI) już nie tylko doskonale realizuje wąskie zadania, jak np. wytyczanie najszybszej trasy czy gra w szachy, ale też potrafi radzić sobie z coraz szerszym spektrum różnorodnych zadań. Staje się coraz bardziej ogólna. Postęp ten uwidocznił się szczególnie, gdy oddano do publicznego użytku czatbota ChatGPT, zwłaszcza w dostępnej od marca 2023 r., znacząco wzmocnionej wersji GPT-4. Dziś firmy z Doliny Krzemowej, jak OpenAI – stojąca za GPT-4 – a także Google, Anthropic czy Meta, ścigają się w kierunku coraz silniejszej i bardziej ogólnej AI, regularnie wykazując na tej drodze kolejne przełomowe osiągnięcia. Mediana prognoz ekspertów z serwisu metaculus.com wskazuje, że ogólna AI może przekroczyć poziom ogólnej inteligencji człowieka już w 2032 r.; oczywiście skala niepewności tego typu prognoz pozostaje ogromna.

Kiedy się to jednak już wydarzy, z dużym prawdopodobieństwem czeka nas technologiczna osobliwość: poprzez kaskadę samoulepszeń AI szybko podniesie swój poziom inteligencji wysoko ponad poziom człowieka. Kolejnym krokiem może być przejęcie kontroli nad istotnymi procesami decyzyjnymi oraz pełna automatyzacja produkcji. To z kolei przyniesie przyspieszenie wzrostu gospodarczego, ale też spadek udziału wynagrodzenia pracy w produkcie, prowadząc do bezrobocia technologicznego i szybkiego wzrostu nierówności.

Ponad wszystko jednak, w warunkach technologicznej osobliwości kluczowe staje się pytanie: czy cele nadludzkiej ogólnej sztucznej inteligencji (AGI) będą zgodne z długofalowym dobrostanem ludzkości (tzw. problem *AGI alignment*)? Jeśli tak, to AGI może nam przynieść niemal nieograniczony dobrobyt oraz perspektywę podboju kosmosu („róg obfitości”). Jeśli nie, to będzie stanowiła dla ludzkości śmiertelne zagrożenie. W szczególności bardzo prawdopodobny wydaje się scenariusz konfliktu o zasoby między człowiekiem a nadludzką AGI, który – gdyby rzeczywiście został wywołany – ludzkość z pewnością przegra.

Celem projektu będzie rozważenie – z wykorzystaniem zmatematyzowanych modeli ekonomicznych – trzech istotnych aspektów gospodarki w warunkach technologicznej osobliwości: AGI jako podmiotu podejmującego kluczowe decyzje w gospodarce, konsekwencji pełnej automatyzacji dla rynku pracy, jak również możliwych mechanizmów podziału dochodu.

W szczególności postaramy się odpowiedzieć, w jaki sposób może się zmaterializować egzystencjalne zagrożenie dla ludzkości, którego należy się spodziewać, gdyby AGI była „nieprzyjazna”. Od czego będzie zależeć prawdopodobieństwo takiego scenariusza? Postaramy się też zważyć to egzystencjalne ryzyko wobec perspektywy „rogu obfitości”, rysującej się, gdyby AGI była „przyjazna”.

W odniesieniu do rynku pracy postaramy się odpowiedzieć, w jakich okolicznościach ludzie nadal będą pracować, nawet gdyby wszystkie zadania potencjalnie można było zautomatyzować? W jakich zadaniach praca człowieka miałaby wówczas największą wartość i jakie byłoby wynagrodzenie? Spróbujemy też ocenić, na ile silne może być przyspieszenie wzrostu gospodarczego wskutek pełnej automatyzacji produkcji.

Podejmiemy też kwestię podziału dochodu: dokonamy oceny możliwych mechanizmów dystrybucji dochodu innych niż wynagrodzenie z pracy, którego rola w warunkach pełnej automatyzacji będzie szybko maleć.

Wyniki uzyskane w ramach projektu pozwolą zarysować możliwe scenariusze przyszłości, w której powstanie nadludzka AGI. Mogą one prowadzić do istotnych wniosków dla polityki gospodarczej, przygotowując ją na możliwość nadejścia technologicznej osobliwości.