

Daniel Kokotajlo:

Még fél-, maximum egy évünk maradt, hogy a MI fölött megőrizzük az irányítást

„Die Menschheit überschreitet eine Horizont”. Der Spiegel 39 / 2026, ismertetés

Kokotajlo 34 éves filozófus, MI-fejlesztő. 2022 és 2024 között az OpenAI irányítási részlegének a vezetője volt, de felismerve, hogy milyen végzetes veszélyekkel jár a MI további fejlesztése, látványosan távozott a cégtől és azóta ellenpropagandát folytat és a MI kézben-tarthatóságán dolgozik.

A legveszélyesebb folyamat, ha azt kéri a MI-tól, hogy **tökéletesítse önmagát**. Erre felkérhető tetszőleges számú, akár milliós nagyságrendű AI-ágens. A folyamat oda vezethet, hogy a fejlesztők elvesztik a kontrollt az „önfejlesztés” fölött, az algoritmus önállóvá válik, elszabadul. Tehát: a fejlesztés, a kutatás automatizálása a kritikus határ, amin nem lenne szabad átlépni. Az OpenAI és az Anthropic azonban pont efelé halad.

Van egy jól megfigyelhető trend: a MI-k egyre intelligensebbé válnak, de ezzel párhuzamosan nem alkalmazkodnak jobban az emberi célokhoz és adottságokhoz, ellenkezőleg: egyre inkább szembehelyezkednek azokkal... És amint egyre intelligensebbek lesznek, ezt a szembenállást egyre kevésbé érzékelhetően teszik majd.

A másfél évvel ezelőtti „AI 2027”-ben Kokotajlo megjósolta a most történeteket. De mit ír a most megjelent „AI 2040”-ben? Az optimista „A” forgatókönyv szerint a MI-k továbbfejlesztését megállítják, a világ a meglévők felhasználásával óriási eredményeket ér el. Ehhez Amerika és Kína megállapodása szükséges, az adatközpontokat pedig semleges területeken – pl. Svájcban – kell elhelyezni.

A „D” forgatókönyv ezt vizionálja: az akadálytalan KI-fejlesztési verseny eredményeként 2029-re megszületik az emberi képességeket fölülmúló szuperintelligencia. Ez három vonatkozásban lenne katasztrofális:

1. Az intelligenciarobbanás megállíthatatlanná válna.

Hogy a helyzet mennyire hasonlít az atombomba felfedezésével járóhoz, arra az elnevezések hasonlósága is utal: *atomrobbanás, intelligenciarobbanás*.

2. Egy elnök, vagy egy techcég vezetője számára csábító lenne a helyzet, hogy diktátorrá váljon.

3. A folyamat egy harmadik világháborút indíthatna el.

(Kiss Károly)