

Eörs Szathmáry:

How AI is Becoming More Biological

Evolvable AI (eAI), i.e., AI systems whose components, learning rules, and deployment conditions can themselves undergo Darwinian evolution, may soon emerge from current trends in generative, agentic, and embodied AI. We argue that this possibility has been underappreciated in debates on AI safety and existential risk. Here, we ask under what technical and ecological conditions AI becomes evolvable, what kinds of behaviours are then likely to emerge, and how such systems could be governed. Drawing on biological evolution and decades of digital evolution experiments, we distinguish “breeder” scenarios, in which humans impose fitness criteria and control reproduction, from “ecosystem” scenarios, in which selection arises from open environments and control erodes. In the latter, selfish replication reliably gives rise to cheating, parasitism, deception, and manipulation, even in very simple systems.

Evolvable AI agents are expected to behave increasingly like biological organisms in cyberspace. They will not only be evolutionary units, but will metabolize, compete, cooperate, and parasitize as part of a kind of Life 2.0. I shall briefly outline how concepts such as individuality, the genotype-phenotype map, fitness, and niche construction will help analyze these concepts.

Szathmáry Eörs:

Hogyan válik az AI egyre inkább biológiai jellegűvé?

Az evolúciós képességű mesterséges intelligencia (eAI) – azaz azok a mesterséges intelligencia-rendszerek, amelyek komponensei, tanulási szabályai és üzemeltetési feltételei maguk is darwini evolúción mehetnek keresztül – hamarosan megjelenhet a generatív, ügynök alapú és megtestesült mesterséges intelligencia jelenlegi trendjeiből. Álláspontunk szerint ezt a lehetőséget eddig alulértékelték a mesterséges intelligencia biztonságával és az egzisztenciális kockázattal kapcsolatos vitákban. Jelen tanulmányban azt vizsgáljuk, hogy milyen technikai és ökológiai feltételek mellett válik a mesterséges intelligencia evolúciós képessé, milyen viselkedésformák alakulhatnak ki ilyenkor, és hogyan lehetne szabályozni ezeket a rendszereket. A biológiai evolúcióra és az évtizedek óta folyó digitális evolúciós kísérletekre támaszkodva megkülönböztetjük a „tenyésztői” forgatókönyveket – amelyekben az emberek szabnak meg alkalmassági kritériumokat és szabályozzák a szaporodást – az „ökoszisztémás” forgatókönyvektől, amelyekben a szelekció nyitott környezetekből fakad, és a kontroll gyengül. Utóbbi esetben az önző replikáció még nagyon egyszerű rendszerekben is megbízhatóan vezet csaláshoz, parazitizmushoz, megfélemlítéshez és manipulációhoz. Az evolúciós képességű mesterséges intelligencia-ügynökök várhatóan egyre inkább biológiai organizmusokhoz hasonlóan fognak viselkedni a kibertérben. Nem csupán

evolúciós egységek lesznek, hanem egyfajta „Life 2.0” részeként anyagcserét folytatnak, versenyeznek, együttműködnek és parazitálnak. Röviden felvázolom, hogyan segítenek az egyediség, a genotípus–fenotípus leképezés, a rátermettség és a niche-építés fogalmai ezeknek a jelenségeknek az elemzésében.