

Mark Coeckelbergh: Etika umělé inteligence

Z anglického originálu „AI Ethics“ přeložila Sylva Ficová.

Praha, Filosofia 2023. 268 s.

Vďaka technologickému pokroku a prepojenosti internetu sa umelá inteligencia stala jedným z hlavných spoločenských fenoménov súčasnosti. Široko diskutovanou témou sú najmä obavy o etickú stránku umelej inteligencie. Touto problematikou sa zaoberá aj kniha Marka Coeckelbergha *Etika umělé inteligence*, ktorú v roku 2023 vydalo nakladateľstvo Filosofia v českom preklade Sylvy Fi-

cové. Pôvodné vydanie knihy vyšlo roku 2020 v rámci série „MIT Press Essential Knowledge“, ktorá sa sústreďí na knihy popredných mysliteľov, ktoré poskytujú odborný prehľad o aktuálnych témach pre neodborníkov. České vydanie knihy je súčasťou edície „Dnešní svět“, ktorá sprístupňuje preklady diel súčasných svetových filozofov o celospoločenských témach širokému okruhu čitateľov. Recenzovaná publikácia si zároveň zachováva grafické spracovanie textu a praktický vreckový formát originálu.

Autor knihy, belgický filozof Mark Coeckelbergh, patrí medzi popredných odborníkov v oblasti filozofie technológií a etiky umelej inteligencie. Je profesorom filozofie médií a technológií na Viedenskej univerzite. Od roku 2023 pôsobí aj ako vedúci výskumný pracovník Centra pre environmentálnu a technologickú etiku pri Filozofickom ústave Akadémie vied Českej republiky. Mark Coeckelbergh je mimoriadne plodný autor, aktívny člen viacerých medzinárodných komisií a expertných skupín zaoberajúcich sa reguláciou umelej inteligencie a uznávaný odborník známy širšej odbornej verejnosti.

Kniha je rozvrhnutá do dvanástich kapitol, ktoré možno tematicky rozdeliť do piatich častí. Prvá časť (kapitoly 1 až 4) sa venuje kultúrnym, historickým a filozofickým súvislostiam spojených s umelou inteligenciou (ďalej AI). Pokrýva mnohé témy: od vízií superinteligencie až po vzťah medzi ľudskou inteligenciou a AI a otázkou morálneho statusu AI. Druhá časť (kapitoly 5 a 6) skúma technickú podstatu AI, jej prepojenie so strojovým učením a dátovou vedou. Tretia časť (kapitoly 7 až 9) analyzuje kľúčové etické problémy AI, ako sú ochrana súkromia, manipulácia, zodpovednosť, vysvetliteľnosť, predpojatosť, spravodlivosť či dopad automatizácie na spoločnosť. Štvrtá časť (kapitoly 10 a 11) sa zameriava na praktické výzvy spojené s implementáciou etických princípov do regulácii AI. Piata, záverečná časť (kapitola 12) kladie etické problémy AI do kontrastu s naliehavosťou iných globálnych problémov a kritizuje iluzórne techno-riešenia.

Coeckelbergh začína svoj výklad poukázaním na to, že AI preniká do každodenných oblastí života, čo vyvoláva nadšenie aj obavy a prináša etické výzvy, ktoré vyžadujú kritickú diskusiu. Úvodnou výzvou, ktorou sa autor zaoberá v druhej kapitole, je myšlienka superinteligencie ako existenčného rizika pre ľudstvo. Túto obavu však považuje za neopodstatnenú a futuristické vízie technologickej singularity a transhumanizmu zasadzuje do širších kultúrnych naratívov, ako sú strach z apokalypsy, frankensteinovský komplex, honba za snom o nesmrteľnosti a gnostické túžby po oslobodení od materiálneho sveta. Varuje, že tieto predstavy môžu skresľovať naše vnímanie AI a odvádzať pozornosť od jej skutočných rizík.

V tretej kapitole si autor kladie otázku, či je všeobecná AI možná. Do hľadania odpovede zapája významných filozofov, napríklad Huberta Dreyfusa, Paula Churchlanda, Daniela Dennetta a Johna Searla, avšak neponúka jednoznačné vlastné stanovisko. Akceptuje, že teoretická možnosť všeobecnej AI existuje, ale jej praktická realizácia je sporná a vzdialená. Následne diskusiu zasadzuje do kon-

textu „dichotómii“: *osvietenstvo vs. romantizmus, humanizmus vs. transhumanizmus a humanizmus vs. posthumanizmus*, ktoré považuje za dôležité pre pochopenie vzťahu medzi človekom a AI. Coeckelbergh sa prikláňa k posthumanistickej vízii, ktorá presadzuje prekročenie ontologických hraníc medzi ľuďmi a neľudskými bytosťami. Táto perspektíva ukazuje, že AI sa nám nemusí a ani nemá podobat', či už v otázke kreativity, alebo inteligencie. V duchu postfenomenológie autor vyzýva, aby sme namiesto obáv z AI upriamili pozornosť na náš vzťah ku konkrétnym technológiám a proces vzájomného spoluvytvárania.

V štvrtjej kapitole sa Coeckelbergh dotýka problému morálneho statusu AI. Odmieťa predstavu AI ako plnohodnotného morálneho aktéra. Pripúšťa však, že AI by mohla mať etické mechanizmy zohľadňujúce dôsledky jej činností. Coeckelbergh sa opatrne prikláňa ku konceptu „funkčnej morálky“, podľa ktorého by určité aplikácie AI, ako napr. autonómne vozidlá, mali byť navrhnuté tak, aby v kritických situáciách dokázali prijať morálne akceptovateľné rozhodnutia. Autor taktiež skúma, či by AI mohla byť objektom morálneho zaobchádzania, voči ktorému by sme mali určité povinnosti, podobne ako voči zvieratám. Hoci sa domnieva, že k aplikáciám AI treba stále pristupovať ako k veciam, naznačuje, že naše chápanie tejto technológie nemusí byť nemenné. Podobne ako sa vyvíjalo naše vnímanie morálneho statusu zvierat, bude sa postupne vyvíjať aj náš postoj k AI. Tento posun bude závisieť od toho, ako hlboko sa AI začlení do našich spoločenských vzťahov.

Prvá časť knihy patrí k filozoficky najzaujímavejším a ponúka komplexný pohľad na AI v kontexte dejín ideí. Coeckelbergh ukazuje, že obavy zo superinteligencie nie sú nové ani eticky relevantné, ale majú korene v tradičných kultúrnych naratívoch. Namiesto antropocentrického pohľadu na technológie navrhuje posthumanistickú perspektívu a zdôrazňuje spoluprácu a vzájomné formovanie ľudí a technológií. Morálny status AI vníma ako dynamický a vzťahový, čo podľa neho otvára možnosť hlbšieho pochopenia a zodpovednejšej integrácie AI do spoločnosti.

V druhej časti knihy sa Coeckelbergh dostáva k technickej stránke AI. V piatej kapitole predstavuje AI ako vedeckú disciplínu aj technológiu a rozoberá jej históriu, technologické aspekty a aktuálne aplikácie. Zdôrazňuje, že komplexnosť AI je kľúčová pre pochopenie jej možností, obmedzení a etických otázok. V šiestej kapitole sa venuje princípom strojového učenia a významu dátovej vedy. Poukazuje na to, ako sa štatistika, pôvodne nudná disciplína zameraná na vyhľadávanie vzorcov a vzťahov v údajoch, stala základom moderných algoritmov strojového učenia. Tieto algoritmy dokážu autonómne analyzovať veľké množstvá dát, rozpoznávať vzory a robiť predpovede. Coeckelbergh však upozorňuje, že napriek ich pokročilosti všetky fázy procesu, od získavania údajov po interpretáciu výsledkov, vyžadujú ľudský zásah.

V tretej časti knihy sa Coeckelbergh zaoberá tradičnými aj menej obvyklými etickými problémami AI. V siedmej kapitole skúma otázky ochrany súkromia, manipulácie používateľov, vplyvu na zraniteľné skupiny, šírenia dezinformácií, politickej manipulácie, bezpečnostných rizík autonómnych vozidiel a zbraní či hrozieb hacknutia. Upozorňuje na koncentráciu moci veľkých spoločností a kritizuje ich za to, že používateľov sociálnych médií využívajú ako neplatenú pracovnú silu na získavanie dát pre komerčné a politické záujmy.

Ôsma kapitola sa venuje problému pripísania morálnej zodpovednosti za škody spôsobené AI a odhaľuje, prečo je tento proces zložitý. Coeckelbergh poukazuje na tradičné podmienky morálnej zodpovednosti, ako aj na špecifické faktory, ako sú rýchlosť rozhodovania, história algoritmov, či problémy „mnohých rúk“ a „mnohých vecí“, ktoré komplikujú určenie zodpovednosti. Zásadnú úlohu, podľa Coeckelbergha, hrá požiadavka vysvetliteľnosti správania strojov, ktorá ovplyvňuje dôveru v technológie. Bez schopnosti porozumieť, prečo AI dospieva k svojim rozhodnutiam, je problematické využívať ju správnym a zodpovedným spôsobom. Coeckelbergh zároveň upozorňuje, že problém vysvetliteľnosti nie je len technickou, ale aj spoločenskou a filozofickou výzvou: Ako môžu bežní ľudia porozumieť týmto vysvetleniam a čo považujeme za adekvátne filozofické vysvetlenie správania AI?

V nasledujúcej kapitole Coeckelbergh skúma dva problémy. Prvým je predpojatosť (bias) v rozhodovaní AI založenej na strojovom učení. Predpojatosť dáť a algoritmov môže prehľbovať existujúce nespravodlivosti a diskriminovať jednotlivcov aj celých skupín. Dáta však nikdy nie sú neutrálne a podľa Coeckelbergha diskriminácia sama o sebe nie je problém – problém nastáva, ak je nespravodlivá. Problém predpojatosti je tak v podstate otázkou spravodlivosti algoritmov. Týka sa rozhodnutia, či má AI zrkadliť svet a ignorovať existujúce rozdiely, alebo pracovať s idealizovanými dátovými súbormi a zahŕňať pozitívnu diskrimináciu. Nejde teda len o technický problém, ale aj o to, akú budúcnosť chceme vytvoriť a aké hodnoty chceme zakomponovať do technológií, ktoré formujú náš svet.

Druhým problémom sú dôsledky automatizácie na budúcnosť práce, spoločenské štruktúry a zmysel života. Coeckelbergh navrhuje prehodnotiť úlohu práce ako zdroja príjmu a zmyslu života. Sympatizuje s myšlienkou všeobecného základného príjmu a podporuje finančné odmeňovanie sociálne dôležitých činností (opatovníctvo a pod.). Zdôrazňuje potrebu spravodlivej integrácie AI do spoločnosti a zamyslenia sa nad významom dobrého života v technologickvej ére. Zároveň kritizuje slabú informovanosť verejnosti a politikov, pričom varuje pred rizikami technokracie, plutokracie a absencie regulácie AI.

Autorov prístup v tretej časti predovšetkým zdôrazňuje zložitosť problematiky a potrebu interdisciplinárneho prístupu pri hľadaní odpovedí na etické otázky AI. K diskutovaným otázkam nezaujíma jednoznačné stanoviská. Namiesto toho ponúka komplexný pohľad, obhajuje rešpektovanie plurality hodnotových systémov

a vyzdvihuje význam demokratického rozhodovania pri navrhovaní algoritmov AI. Hoci naznačuje potrebu vytvorenia teórie spravodlivosti AI, bližšie nešpecifikuje, ako by takáto teória mohla vyzerat'.

Desiata a jedenásta kapitola tvoria štvrtú časť knihy, venovanú problémom tvorby a implementácie regulačných návrhov. Coeckelbergh poskytuje prehľad existujúcich návrhov v oblasti etiky AI, vrátane politických, akademických, firemných, občianskych a profesijných iniciatív. Konštatuje, že síce existuje konsenzus ohľadom hlavných princípov (napr. vysvetliteľnosť, transparentnosť), ich praktická realizácia však zostáva nejasná.

V nasledujúcej kapitole upozorňuje na potrebu iniciatívnej etiky – integrácie etiky do fázy návrhu a vývoja. Tento prístup však čelí viacerým obmedzeniam, ako sú neistota ohľadom budúcich dôsledkov nových technológií, nedostatočná participácia zainteresovaných strán, koncentrácia moci v rukách technologických korporácií a ťažkosti s prekladom kultúrne podmienených hodnôt do technologického dizajnu. Coeckelbergh vyzýva na demokratickejší a inkluzívnejší prístup k etike AI, ktorá by nemala byť len sériou pravidiel, ale komplexnou filozofickou víziou (pozitívna etika), ktorá usmerňuje technologický pokrok v prospech všetkých.

V štvrtej časti autor čerpá zo svojich praktických skúseností člena expertných skupín v oblasti etiky AI. Predstavuje kľúčové iniciatívy a problémy pri implementácii etických princípov do regulačných opatrení. Apeluje na interdisciplinárnu spoluprácu a užšie prepojenie technických a humanitných disciplín vo vzdelávaní. Uznáva však, že lepšie vzdelávanie nestačí a konkrétne kroky v konečnom dôsledku závisia od politickej vôle.

V poslednej časti knihy, kapitole dvanásť, prináša Coeckelbergh provokatívny pohľad na etiku AI v kontexte environmentálnych problémov. Zdôrazňuje naliehavosť klimatických výziev a kladie otázku, či by obavy z AI mali byť prioritou. Varuje, že technológie môžu odvádzať pozornosť od reálnych problémov, ako sú klimatická kríza, chudoba či nedostatok pitnej vody, a kritizuje vízie kolonizácie iných planét ako riešenia pre Zem. Upozorňuje tiež na riziká techno-riešení, ktoré môžu podporovať paternalizmus a ohroziť slobodu. V závere zdôrazňuje, že riešenie globálnych problémov si vyžaduje viac než technológiu – je potrebná etická a sociálna reflexia spájajúca inteligenciu s praktickou múdrosťou.

Autor uzatvára svoju diskusiu o etických problémoch AI pomerne neortodoxným spôsobom, keď relativizuje ich význam v kontraste s nevyriešenými globálnymi problémami. Svoju kritiku smeruje tiež k zjednodušujúcim techno-optimistickým riešeniam a vizionárskym predstavám hlásajúcim únik od reality pozemských problémov ľudstva. Autor sa snaží zdôrazniť, že AI by nemala slúžiť ako nástroj odcudzenia, ktorý oddelí ľudí od ich existencie tu a teraz.

Coeckelberghova kniha presahuje tradičné otázky etiky AI a rozoberá široké spektrum tém: od vzťahu AI k náboženstvu, duchovným a filozofickým otázkam cez technické aspekty, sociálnu kritiku, otázku spravodlivosti a dobrého života až

k tvorbe politík a globálne výzvy udržateľnosti. Tomuto interdisciplinárnemu prístupu je podriadená aj štruktúra knihy a jednotlivých kapitol. Hoci kapitoly sledujú premyslenú postupnosť tém, štruktúra niektorých kapitol pôsobí nesúrodno. Autor často diskutuje rovnaké alebo veľmi podobné témy v rôznych kapitolách, čo napriek odlišnému kontextu vytvára dojem opakovania a môže čitateľa miasť.

Kniha je mimoriadne bohatá na otázky, ktoré ponúkajú široký rámec na uvažovanie o problémoch týkajúcich sa etiky AI. Mnohé z otázok sú zámerne otvorené a majú skôr viesť konkrétnu tému alebo podnietiť hlbšie úvahy. Avšak pre čitateľa môže byť po čase frustrujúce, že autor kladie množstvo otázok, na ktoré vlastne neponúka odpovede. Text tak občas zanecháva pocit neúplnosti a nedostatku smerovania.

Iné otázky a problémy autor síce rozvinie, ale ponúkne skôr sumarizáciu názorov než jednoznačné odpovede. V otázke všeobecnej AI Coeckelbergh naznačuje filozofické a technické argumenty pre a proti, ale v podstate ju necháva otvorenú. Upozorňuje na potrebu spravodlivej AI, no problém skôr kontextualizuje, než že by prinášal riešenia. Vo viacerých kapitolách sa zmieňuje o koncentrácii moci v rukách technologických gigantov a potrebe regulácie, ale riešenie tejto otázky ponecháva bez konkrétnych návrhov. Hoci vyzýva na integráciu etiky do vývoja AI, ide skôr o výzvu na dôveru v etiku než o praktické riešenia. Treba však poznamenať, že mnohé otázky, ktoré Coeckelbergh kladie, v súčasnosti buď nemajú jasné odpovede, alebo medzi odborníkmi neexistuje zhoda na ich riešení. Otázky položené v knihe tak len odrážajú nejednoznačnosť a pluralitu názorov na dané témy v súčasnej odbornej diskusii.

Azda najvýraznejším prínosom knihy je Coeckelberghova schopnosť poukázať na to, že výskum AI neslúži len na technologický pokrok, ale zároveň otvára aj dôležité otázky o nás samých – o našej prirodzenosti, hodnotách a spôsobe, akým organizujeme spoločnosť. Autor umne zdôrazňuje reflexívny rozmer AI a upozorňuje na to, že snaha o konštrukciu inteligentných systémov nás vedie k lepšiemu pochopeniu fungovania ľudskej inteligencie, myslenia, učenia, kreativity a rozhodovania. Tento proces odhaľuje naše predsudky, predpojatosti, spoločenské nerovnosti a hodnoty, ktoré v spoločnosti preferujeme. Coeckelbergh tiež upozorňuje na nedostatok odborných znalostí a nepripravenosti na rýchly technologický vývoj a výzvy spojené s koncentráciou technologickéj moci. Ponúka tak perspektívu, ktorá presahuje tradičný rámec a prináša hlbšie filozofické, antropologické, sociálne, právne a politické implikácie AI, čím výrazne obohacuje diskusiu o etike AI.

Ivan Koniar

Filozofická fakulta Katolíckej univerzity v Ružomberku
ivan.koniar@ku.sk