Hodnoty a etické výzvy využívání systémů umělé inteligence ve společnosti

MUDr. Anetta Jedličková, Ph.D.

Fakulta humanitních studií, Univerzita Karlova Praha

Pracoviště doktorských studií, Obor Aplikovaná etika

Pátkova 2137/5, 182 00 Praha 8

email: Anetta.Jedlickova@fhs.cuni.cz

Values and ethical challenges of using artificial intelligence systems in society

Abstract

Technologies that use artificial intelligence in their procedures are increasingly becoming an indispensable part of our work, as well as personal activities in many areas. The unprecedented development of autonomous and intelligent systems has arisen numerous new challenges and various ethical aspects or ethical dilemmas that already need to be pointed out when designing an autonomous system using the machine learning technology in order to prevent any serious social consequences. Trustworthy artificial intelligence systems should be developed, deployed and used in a safe, transparent and accountable manner, in compliance with ethical principles and with respect for the moral values, such as humanity and human dignity, the rights to respect for privacy and protection of personal data, justice and fairness. Artificial intelligence systems should increase benefits and utilities that are shared by all society. The paper summarizes the potential aspects that must be taken into account in the field of artificial intelligence, and acquaints professionals, as well as the general public with some ethical implications that may occur during the development, implementation and use of artificial intelligence, robotics and related technologies for the material and immaterial integrity of individuals or groups, as well as for the society as a whole.

Key words

artificial consciousness, artificial intelligence, autonomous and intelligent systems, autonomous decision-making, ethical challenges, ethical dilemmas, ethical principles, moral values

Úvod

Technologie, které využívají při svých postupech umělou inteligenci, se stále častěji stávají nepostradatelnou součástí pracovních i osobních činností člověka v mnoha oblastech. Nebývalý rozvoj autonomních a inteligentních systémů s sebou přináší rozličné etické aspekty, které mohou ovlivnit společnost, společenské blaho a obecné dobro. Systémy umělé inteligence nabízejí na jedné straně značné příležitosti a výhody, na druhé straně však vyvolávají určitá rizika, kterými se je potřeba náležitě zabývat, protože jejich dopady představují aktuální společenský problém v mezinárodním měřítku.

Podstatným úkolem je zajištění důvěry v základní cíle využívání systémů umělé inteligence ve společnosti, k nimž patří nejen posílení individuálního a společenského blaha a obecného dobra, ale také zajištění pokroku a inovací a usnadnění dosažení cílů udržitelného rozvoje stanovených Organizací spojených národů, jako jsou boj proti změně klimatu, hospodárné využívání přírodních zdrojů, zlepšení zdraví, mobility a výrobních procesů, podpora způsobu sledování pokroku na základě ukazatelů udržitelnosti a sociální soudržnosti či podpora genderové vyváženosti. Systémy založené na umělé inteligenci musí být vždy zaměřené na dobro pro člověka, musí usilovat o maximalizaci přínosů a o zavedení náležitých opatření s cílem rizika minimalizovat. Důvěryhodné systémy umělé inteligence musí být vyvíjeny, zaváděny a používány bezpečně, transparentně a zodpovědně, v souladu s etickými principy, s respektem k lidské důstojnosti, s právem na respektování soukromí a ochranu osobních údajů. Závazek důvěryhodných systému v oblasti umělé inteligence se týká všech zúčastněných stran, které navrhují, vyvíjejí, zavádějí, implementují nebo používají systémy umělé inteligence nebo jsou jí dotčeny, a to jak jednotlivců, pracovníků, spotřebitelů či výzkumných pracovníků, tak podniků, organizací, veřejných služeb, vládních orgánů či institucí (High-Level Expert Group 2019).

Přestože v českém a slovenském prostoru studie a výzkumy využívající systémy umělé inteligence probíhají, v odborných časopisech však ucelený souhrn společensky důležitých požadavků na jejich návrh, vývoj, implementaci a používání absentuje. Cílem tohoto článku je

proto poskytnout přehled současného stavu vývoje etických požadavků v dané oblasti a poukázat na etické aspekty, které mohou mít zásadní etické implikace na uživatele a celou společnost. Text je rozdělen na několik částí. Nejprve je stručně vysvětlen pojem umělé inteligence, její základní dělení a využití. Následně jsou představena teoretická východiska, včetně etických principů. Na ně navazuje výčet etických aspektů a potenciálních dopadů, které jsou při využívání systémů umělé inteligence významné, což nás přivede k roli základních morálních hodnot při jejich vývoji, implementaci a využívání. V závěru je představena konkrétní studie, která v dané oblasti aktuálně probíhá.

Základní dělení umělé inteligence a její využití

Pod pojmem umělá inteligence rozumíme systémy vykazující inteligentní chování na základě zpracovávání a analýzy prostředí či poskytnutých dat a následného rozhodování, přijímaní a provádění opatření s určitou mírou autonomie k dosažení konkrétních cílů (Evropská komise 2018). Pro umělou inteligenci je i v našich podmínkách často používán její anglický ekvivalent Artificial Intelligence (AI), případně také pojem autonomní a inteligentní systémy (anglicky autonomous and intelligent systems – A/IS, případně taktéž AI/IS), který ve svých dokumentech používá například Institute of Electrical and Electronics Engineers (IEEE).

Umělou inteligenci lze rozdělit na tři základní skupiny podle úrovně jejích schopností nalézt řešení:

1. Artificial narrow intelligence (ANI), tzv. úzká umělá inteligence, má úzký a přesně vymezený rozsah schopností. Zaměřuje se na nejkvalitnější a nejoptimálnější možné splnění specifického úkolu (např. rozpoznávání obličejů, hlasů, roboticky asistovaná chirurgie, parkovací asistent automobilů a další). V současnosti je ANI v různých oblastech běžně využívaná a její využití v praxi se neustále rozšiřuje.
2. Artificial general intelligence (AGI), neboli obecná umělá inteligence, má rozsah schopností na stejné úrovni jako jsou schopnosti přirozené pro člověka. Jejím cílem je využívání inteligence takovým způsobem, aby byla stejně kompetentní, jako je lidská mysl, a tedy realizování úkonů na úrovni lidské inteligence. Vývojem AGI se zabývá řada výzkumných projektů, v současnosti však neexistuje žádný funkční systém, který by byl svými rekurzivními algoritmy vylepšován na úroveň lidské mysli.
3. Artificial superintelligence (ASI), tzv. super-inteligence, jde o pojem pro dosud neexistující úroveň inteligence, která by výrazně přesahovala schopnosti člověka a lidskou

inteligenci by převyšovala v různých kognitivních oblastech. ASI lze dále teoreticky dělit na rychlostní, kolektivní a kvalitativní super-inteligenci podle způsobu překonávání schopností člověka, například v rychlosti, výkonu či kvalitě kognitivní lidské inteligence (Bostrom 2014).

Technologie využívající umělou inteligenci disponují schopnostmi zcela autonomního přizpůsobování se rozličným změnám či předvídání dalšího vývoje určitého stavu, a na základě vyhodnocení dat se tak mohou autonomně rozhodovat pro optimální řešení dané situace, a tím přinášet prospěch jedinci i společnosti. V některých situacích však přinášejí také riziko, že svým autonomním rozhodováním mohou ohrozit, či dokonce způsobit újmu v různých oblastech společnosti, proto je nutné podporovat rozvoj sociálně odpovědné umělé inteligence, důsledně kontrolovat směr jejího vývoje a související rizika a včas přijímat odpovídající opatření k zamezení jakéhokoli negativního dopadu, poškození či újmy.

Etické principy rozvoje umělé inteligence, robotiky a souvisejících technologií

Etické aspekty algoritmického rozhodování a strojového učení při zpracování a analýze dat patří k významným tématům, kterými se odborná veřejnost zabývá v souvislosti s intenzivním vývojem technologií využívajících umělou inteligenci. Každodenní praxe s sebou přináší etické požadavky na zajištění bezpečnosti jednotlivců i společnosti. Je nutné podotknout, že umělá inteligence při svém autonomním rozhodování nečiní žádná morální rozhodnutí, a tedy nelze zajistit její etické jednání či rozhodování ve smyslu prospěšnosti a maximalizace přínosů pro jedince a společnost, neškodlivosti a nezhoršování jakékoli újmy, ochrany nedotknutelnosti, minimalizace rizik, zajištění spravedlnosti, nediskriminace či prevence stigmatizace, jakož i respektu k autonomii uživatelů a všech zúčastněných, respektování jejich svobodné vůle a rozhodnutí s vyloučením jakéhokoli ovlivňování či manipulace, s čímž je úzce spojeno právo na lidskou důstojnost a svobodu. Bez adekvátní lidské kontroly, a to již během navrhování autonomních a inteligentních systémů a následně také během jejich vývoje, zavádění a používání, může z těchto důvodů docházet k závažným etickým implikacím.

Významnou roli při zajištění ochrany práv, soukromí, lidské důstojnosti či svobody všech zúčastněných má dodržování čtyř základních etických principů[1], které jsou v praxi již běžně aplikovány v medicíně a v biomedicínském výzkumu, tj. princip beneficence, princip nonmaleficence, princip spravedlnosti a princip respektu k autonomii člověka (Beauchamp, Childress 2019). Princip beneficence apeluje na využívání technologií pouze ve prospěch jedince a společnosti a na maximalizaci přínosů pro jedince a společnost. Princip nonmaleficence vyžaduje minimalizaci rizik, vyhodnocování poměru risk/benefit, předcházení vzniku individuální a kolektivní újmy, včetně nehmotné újmy (morální, sociální, společenské či psychické újmy) a ochranu fyzické a duševní nedotknutelnosti. Princip spravedlnosti spočívá ve spravedlivém rozdělení přínosů a rizik pro všechny zúčastněné strany, v zajištění spravedlivého přístupu, stejných práv, rovných příležitostí, rovnosti podmínek při rozhodování, v zamezení nespravedlivé podjatosti, diskriminace a stigmatizace a také v dodržování rovnováhy mezi protichůdnými zájmy a cíli či v možnosti zpochybnit nespravedlivá rozhodnutí přijatá autonomním systémem a domoci se jejich účinné nápravy. Princip respektu k autonomii člověka je založen na respektování svobodné vůle uživatelů a jejich rozhodnutí, na vyloučení jejich ovlivňování či manipulace, umožnění jednotlivcům činit odůvodněná informovaná rozhodnutí, na respektu k jejich hodnotovému systému, na ochraně fyzických osob v souvislosti se zpracováním osobních údajů, volným pohybem těchto údajů a jejich správou a do této skupiny patří také problematika mlčenlivosti.

Odborná skupina High-Level Expert Group on Artificial Intelligence (AI HLEG), zřízena Evropskou komisí v červnu 2018, přidala ve svých etických pokynech pro důvěryhodnou umělou inteligenci *(Ethics guidelines for trustworthy AI)* další etický princip, který výše uvedené čtyři základní etické principy v souvislosti s moderními technologiemi vhodně doplňuje. Jde o princip vysvětlitelnosti prostřednictvím transparentnosti, srozumitelnosti a odpovědnosti (High-Level Expert Group 2019). Každý z uvedených principů obsahuje v souvislosti s autonomními a inteligentními systémy konkrétní požadavky, které je nutné dodržovat ve všech fázích života autonomních a inteligentních systémů.

V říjnu 2020 byl Evropským parlamentem schválen regulační rámec pro umělou inteligenci s názvem *Rámec pro etické aspekty umělé inteligence, robotiky a souvisejících technologií (Framework of ethical aspects of artificial intelligence, robotics and related*

[1] V souvislosti s etickým rozhodováním v medicíně vydali Tom L. Beauchamp a James F. Childress v roce 1979 první vydání Principles of Biomedical Ethics. Aplikovaný etický přístup k řešení etických dilemat založen na aplikaci určitých etických principů se označuje anglickým termínem Principlism a lze jej využít v jakékoli oblasti aplikované etiky, nikoli pouze v bioetice.

technologies), který kromě právních povinností specifikuje rovněž podstatné etické zásady pro vývoj, zavádění a používání umělé inteligence, robotiky a souvisejících technologií a připomíná důležitost prevence jakýchkoli rizik, která by mohla ohrozit bezpečnost jednotlivců i společnosti. Regulační rámec dále připomíná povinnost plně dodržovat požadavky zakotvené v *Listině základních práv Evropské unie* (dále jen Listina), zejména zachovávání lidské důstojnosti, autonomie a sebeurčení jednotlivce (Evropský parlament a Rada EU 2012), dále povinnost předcházet škodám, prosazovat spravedlnost, začleňování a transparentnost, odstraňovat předsudky a diskriminaci, a to také v souvislosti s menšinovými skupinami, respektovat a dodržovat zásady spočívající v omezování negativních vlivů používaných technologií, zajišťování jejich vysvětlitelnosti a zaručování požadavku, aby technologie umělé inteligence sloužily lidem, aby je však nenahrazovaly, případně za ně nerozhodovaly.
V dokumentu je zdůrazněno, že vývoj, zavádění a používání umělé inteligence, robotiky a souvisejících technologií musí usilovat o zvyšování blahobytu a svobody jednotlivců, o zachování míru, předcházení konfliktům, zvyšování mezinárodní bezpečnosti a zároveň musí maximalizovat poskytované přínosy, předcházet souvisejícím rizikům umělé inteligence a snižovat je. V neposlední řadě obsahuje regulační rámec také požadavky na umožnění zapojení lidského faktoru pro nezávislý dohled prostřednictvím vhodných kontrolních opatření a převzetí kontroly odpovědnou a kvalifikovanou osobou v případě potřeby (European Parliament 2020).

Etické konotace vývoje, zavádění a používání umělé inteligence

Rozvoj moderních technologií využívajících umělou inteligenci, jež jsou schopny autonomně rozhodovat, s sebou přináší nespočet etických aspektů, které je nutné zohledňovat již při vývoji, následném zavádění i samotném používaní autonomních systémů. Umělá inteligence, robotika a související technologie, včetně softwaru, algoritmů a dat, které tyto technologie používají či produkují, a to bez ohledu na oblast, v níž jsou vyvíjeny, zaváděny nebo používány, by se měly již od svého návrhu vyvíjet bezpečným, transparentním, technicky detailně propracovaným, spolehlivým a právně závazným způsobem. Jedná se o uplatňování zásady „etika již od návrhu“. Na tomto místě se budeme věnovat některým závažným oblastem, které mají na etické dopady využívaní technologií umělé inteligence ve společnosti významný vliv.

1. Důvěra uživatelů

Pro zajištění důvěry společnosti a jednotlivců jsou zásadní konkrétní regulační předpisy a pokyny týkající se vysvětlitelnosti, auditovatelnosti, sledovatelnosti a transparentnosti. Důvěra uživatelů má podstatný význam pro další rozvoj a uplatňování moderních technologií v praxi, proto by spotřebitelé měli mít právo být náležitě informováni, a to srozumitelným, včasným, standardizovaným, přesným a přístupným způsobem, nejen o existenci a náležitém odůvodnění určitého autonomního systému, ale také o možných výsledcích a praktických dopadech používaných algoritmických systémů. Poskytnutí informací o způsobech potenciální obrany spotřebitele a o možnostech kontaktování odpovědné osoby s rozhodovacími pravomocemi by mělo patřit k základním pravidlům. Rozhodnutí autonomního systému s významným dopadem na jednotlivce či společnost musí kromě možnosti kontroly nabízet taktéž možnost smysluplného zpochybnění takového rozhodnutí a v případě potřeby také možnost opravy. Každá fyzická i právnická osoba by měla mít možnost domáhat se nápravy rozhodnutí, které vydala technologie umělé inteligence, robotika nebo související technologie v její neprospěch nebo v rozporu s právem. Z uvedeného vyplývá, že je důležité zajistit účinné nápravné prostředky, tedy přístupné, cenově dostupné a nezávislé postupy a mechanismy přezkumu, které by zaručovaly nestranné lidské posouzení případného porušení práv, jako jsou práva spotřebitelů nebo občanská práva, nezávisle na sektoru, v němž jsou příslušné algoritmické systémy používány (tj. veřejný, komerční či soukromý sektor). V této oblasti je nutné připomenout také poskytování náležité pomoci společnosti při uplatňování práva spotřebitelů na nápravu v případech porušení jejich práv. Důvěra na všech úrovních zúčastněných stran a společnosti v umělou inteligenci, robotiku a související technologie, zejména jsou-li považovány za vysoce rizikové, vyžaduje na jedné straně omezení úřednické zátěže a byrokracie a zároveň vytvoření efektivního legislativního rámce v souladu s etickými zásadami s cílem podpořit právní jistotu a zaručit základní práva a ochranu spotřebitele.

K vytvoření a posílení důvěry spotřebitelů je neméně důležitá řádná ochrana sítí propojené umělé inteligence a robotiky a přijetí přísných opatření s cílem předejít případům narušení bezpečnosti, úniku dat, tzv. „otrávení dat“ (data poisoning), kybernetickým útokům a zneužívání osobních údajů. Významná rizika představují především autonomní systémy, které jsou založeny na neobjektivních datových souborech a neprůhledných algoritmech. Proto je zásadní, aby všichni aktéři v celém vývojovém i dodavatelském řetězci produktů a služeb

umělé inteligence (vývojáři, provozovatelé a uživatelé) nesli jasně vymezenou právní odpovědnost za případné újmy (European Parliament 2020).

2. Nezaujatost a nediskriminace

Jak je upozorněno ve výše uvedeném regulačním rámci pro etické aspekty, umělá inteligence má v závislosti na způsobech svého vývoje a používání významný potenciál vytvářet a posilovat předsudky, a to také prostřednictvím inherentních předsudků v základních datových souborech, a vytvářet tak různé formy automatizované diskriminace včetně nepřímé diskriminace, které se týkají zejména skupin lidí s podobnými charakteristikami.

Technologie umělé inteligence mohou případnými předsudky, diskriminací nebo stigmatizací prostřednictvím vytvořeného softwaru, algoritmů či dat způsobit jednotlivcům i společnosti zjevnou újmu. Může se jednat o újmu fyzickou, psychickou, materiální nebo také nehmotnou újmu. Zajištění objektivity a kvality vstupních datových souborů a vytvoření pravidel pro zpracovávání údajů takovým způsobem, aby autonomní technologie neprodukovaly neobjektivní výstupy, tvoří zásadní parametry k zamezení rizika předpojatosti a diskriminace při vývoji autonomních a inteligentních systémů. Technologie umělé inteligence by měly být navrhovány tak, aby dodržovaly a chránily fyzickou a mentální integritu a podporovaly kulturní, jazykovou a individuální rozmanitost. Je nutné přijmout účinná opatření, aby technologie umělé inteligence nebyly používány způsobem, který by mohl vést k nepřípustnému přímému či nepřímému nátlaku, oslabení autonomie, neodůvodněnému sledování, klamání či nepřípustné manipulaci. Základem těchto technologií by měly být hodnoty, jako jsou spravedlnost, respekt k autonomii, důvěrnost a transparentnost, lidská důstojnost a další. Záměrem využívání moderních technologií v boji proti předsudkům a diskriminaci by mělo být zajištění rovnosti příležitostí, genderové rovnosti, řádné zohledňování a zastupování zájmů všech osob, včetně marginalizovaných skupin či osob ve zranitelném postavení (například osoby se zdravotním postižením), a dosažení rovných práv a pozitivních sociální změn (European Parliament 2020).

3. Sociální odpovědnost

Další oblast umělé inteligence, robotiky a souvisejících technologií, která má významný etický dopad na jednotlivce i společnost, je sociální odpovědnost. Sociálně odpovědné technologie založené na umělé inteligenci pomáhají nalézt řešení, která mohou zachovat a podpořit rozličné cíle týkající se společnosti, například řešení na podporu sociálního začleňování, plurality, solidarity, spravedlnosti, rovnosti či spolupráce. Jedná se o řešení, která chrání a podporují základní práva a hodnoty společnosti, jako jsou demokracie, právní

stát, ochrana dětí a zdraví, hospodářská prosperita, pracovní a sociální práva, kvalitní a dostupné vzdělávání, pluralitní a nezávislé sdělovací prostředky, objektivní a volně dostupné informace či digitální gramotnost. Autonomní systémy nesmí úmyslně způsobovat žádnou újmu jednotlivcům, společnosti ani životnímu prostředí v žádné z uvedených oblastí. Jako příklad uveďme finanční nebo hospodářskou ztrátu, ztrátu zaměstnání, ztrátu příležitosti ke vzdělávání, neoprávněné omezení svobody volby nebo svobody projevu, ztrátu soukromí či jakékoli porušení práva.

Technologie, jejichž vývoj, zavádění a používání přinášejí významné riziko způsobení újmy jednotlivcům nebo společnosti, by měly být považovány za vysoce rizikové technologie a mělo by se důsledně zvažovat, v jakém odvětví budou využívány, jejich speciální způsob a účel použití či závažnost újmy, kterou by mohly způsobit. Míra závažnosti by se měla stanovovat na základě míry potenciální újmy, množství újmou postižených osob a celkové hodnoty případné škody a újmy způsobené společnosti jako celku. K závažným druhům újmy patří například porušení práv dětí, spotřebitelů nebo pracovníků, které kvůli svému rozsahu, počtu nebo jejich dopadu na společnost jako celek přinášejí riziko negativního dopadu na fyzickou a duševní pohodu, nenávistné verbální projevy, násilí apod. (European Parliament 2020)

4. Ochrana osobních údajů

Je nezpochybnitelné, že v důsledku vývoje, zavádění a používání umělé inteligence, robotiky a souvisejících technologií se podstatně zvyšuje používání dat v praxi, včetně osobních údajů, jako jsou biometrické údaje. Tyto technologie skýtají potenciál používat osobní a neosobní údaje k následné kategorizaci osob, k odhalení zranitelných míst jednotlivců či k využití údajů na přesné zacílení osob nebo skupiny osob.

Účinně prosazované zásady ochrany údajů a soukromí jsou založeny na důležitých omezeních a kontrolních mechanismech, jako jsou např. minimalizace údajů, právo odmítnout profilování, kontrola použití svých údajů, právo získat vysvětlení rozhodnutí, které je založené na automatizovaném zpracování, ochrana soukromí již od návrhu, omezení na základě předem přesně vymezeného účelu, jakož i zásady proporcionality a nezbytnosti.

Zvláštní rizikovou skupinu představují osobní údaje týkající se nebo pocházející od zranitelných skupin, jako jsou osoby se zdravotním postižením, pacienti, děti, starší osoby, menšiny, migranti a další skupiny ohrožené vyloučením. Jejich ochraně musí být věnována zvláštní pozornost, je eticky nepřijatelné jakékoli zneužití jejich osobních údajů.

Je nutné pečlivě zvážit etické hranice pro používání umělé inteligence, robotiky a souvisejících technologií, včetně softwaru, algoritmů a dat, které tyto technologie používají či produkují, především v situacích, v nichž se využívají technologie na dálkové rozpoznávání, jako je rozpoznávání biometrických charakteristik (zejména rozpoznávání obličeje) pro automatické zjišťování totožnosti osob. Používají-li tyto technologie veřejné orgány z důvodu zásadního veřejného zájmu, například za účelem zajištění bezpečnosti jednotlivců a řešení vnitrostátních krizových situací či zabezpečení bezpečnosti majetku, měly by je používat vždy veřejně, přiměřeně a cíleně, pouze na konkrétní účely, po omezenou dobu a v souladu s právními předpisy, základní právy zakotvenými v Listině a s řádným přihlédnutím k lidské důstojnosti a autonomii. Jak zdůrazňuje regulační rámec pro etické aspekty, kritéria pro tato užití a jejich omezení by měla podléhat soudnímu přezkumu, kontrole a diskusi, do níž bude zapojena občanská společnost (European Parliament 2020). Vždy je zcela nezbytné dodržovat a prosazovat práva občanů na soukromí a ochranu osobních údajů, včetně osobních údajů odvozených od neosobních a biometrických údajů, a to v souladu s relevantními právními předpisy a etickými principy.

5. Kontrola autonomních systémů

Rozhodnutí produkovaná umělou inteligencí, robotikou a souvisejícími technologiemi by měla nadále podléhat přiměřenému lidskému přezkumu, posouzení, zásahům a kontrole. Důvěryhodné autonomní systémy musí být nejen vyvíjeny, zaváděny a používány bezpečně, transparentně a zodpovědně v souladu s bezpečnostními prvky týkajícími se robustnosti, odolnosti, bezpečnosti, přesnosti a identifikace chyb, vysvětlitelnosti, transparentnosti a identifikovatelnosti, ale také náležitě a pravidelně kontrolovány. Technická a provozní složitost autonomních technologií by neměla bránit jejich provozovateli nebo uživateli, aby bylo v každém okamžiku umožněno nouzové odstavení, změna či zastavení jejich provozu nebo navrácení k předchozímu stavu obnovou bezpečnostních funkcí. Kontrola a dohled by měly být zaměřeny na dodržování požadavků na kvalitu, integritu, transparentnost, důvěryhodnost, bezpečnost a ochranu údajů a soukromí v návaznosti na dodržování základních etických principů. V souladu s regulačním rámcem pro etické aspekty by měla být transparentnost zajištěna také umožněním přístupu veřejných orgánů k technologiím, datům a relevantním počítačovým systémům v nezbytně nutných případech (European Parliament 2020).

Etické implikace používání autonomních a inteligentních systémů

Zajištění kybernetické bezpečnosti je stěžejním úkolem všech odpovědných článků účastnících se při vývoji, zavádění a používaní technologií s umělou inteligencí. Otázkou však zůstává, zda a jakým způsobem lze zajistit, aby autonomní a inteligentní systémy striktně sloužily jednotlivci a společnosti, nebyly zneužitelné žádnými mocenskými vlivy a jejich rozhodnutí byla eticky správná. Je zřejmé, že žádná umělá inteligence v současné době neumí posoudit etické dopady svého rozhodnutí, a neumí tedy eticky rozhodovat. Umí se to naučit v současnosti známými technologiemi strojového učení (machine learning)? Jistě by do jisté míry zvládla rozhodovat podle základních etických principů či teorií, které bychom jí poskytli, a uměla by se naučit používat je podle hodnot a paradigmat, jimiž bychom ji vybavili. Zvládla by je však použít správně ve všech situacích, v nichž člověk zvažuje na individuální úrovni různé varianty, které se mohou v návaznosti na různost okolností značně lišit? Bude mít schopnost individuálního posouzení, rozlišení či citu pro konkrétní situaci? K eticky správnému rozhodnutí dospěje člověk konkrétní individuální rozvahou ovlivněnou situačními odlišnostmi, nikoli pouze naučenými mechanismy, postupy, principy, pravidly, normami. Etická rozvaha je konkrétní a vztahuje se pouze na daný moment za daných okolností. Zvládla by umělá inteligence také správně samostatně vyřešit náročná etická dilemata, jejichž zvažování a následné rozhodování je mnohokrát i pro člověka velmi obtížně? Pro ilustraci uveďme příklad lékaře, který podle etického principu beneficence léčí, a tedy pomáhá a usiluje o prospěch a dobro. Podle principu nonmaleficence se snaží minimalizovat riziko zhoršení zdravotního stavu pacienta a nezpůsobit mu svým konáním úmyslnou újmu. Podle dalších principů respektuje preference pacienta, které reflektují jeho vnitřní hodnoty, vůli a cíle (princip respektu k autonomii) a spravedlivě aplikuje dostupnou léčbu či chrání vulnerabilní skupiny pacientů (princip spravedlnosti). V každé konkrétní situaci provádí kromě terapeutické rozvahy také důkladnou kvalifikovanou etickou rozvahu, během níž individuálně posuzuje nejen naplňování čtyř základních etických principů, ale taktéž všechny dostupné informace a zvažuje celou řadu dalších faktorů, jako jsou například cíle terapeutické intervence, kvalita následného života, informace získané ze vzájemné komunikace s pacientem a pochopení jeho hodnot a priorit, a to vše zcela individuálně u každého jednotlivého pacienta (Jedličková 2020). Na čem je založeno vedení konkrétní etické rozvahy u člověka: jde o naučenou rutinu, nebo snad k tomu potřebujeme určité specifické atributy, kterými umělá inteligence nedisponuje? Lidé mají různé psychické i fyzické zkušenosti, představy, nápady, pocity, reflexe i sebereflexe, intuici a empatii, a to vše rozhodování zásadně ovlivňuje. Jde o poznání spojené s vědomým vnímáním všech konkrétních okolností,

ovlivněné nejen racionálním uvažováním, ale také intuitivním a emocionálním uvažováním, s uvědomováním si souvislostí, přesahů a vzájemných propojení či prolínání potenciálních důsledků. Má, resp. může vlastnit také umělá inteligence stejné atributy a má-li předčit schopnosti člověka, mohou být dokonce kvalitnější? A lze-li těmito lidskými kvalitami autonomní a inteligentní systémy obdařit, budou poté při konfrontaci s etickými dilematy již skutečně disponovat schopnostmi vést adekvátní kvalifikovanou etickou rozvahu na zcela individuální úrovni při posouzení všech výše uvedených faktorů, které mimo jiné vyplývají z empatie, z různých individuálních a společenských hodnot a požadavků, ze správného porozumění kontextu, z pokročilé sociální interakce a v neposlední řadě také z porozumění významu a důsledků vlastního zpracování informací a rozhodnutí? Je evidentní, že skutečného morálního uvažování nelze dosáhnout pouze zlepšením výpočetních schopností, algoritmů či strojového učení a je nutné přidat některou nezbytnou součást lidské výbavy.

Důležitost těchto úvah nabývá ještě větší relevance, uvažuje-li se o možnosti vytvořit umělou inteligenci, která by byla schopná disponovat takovým souborem charakteristik či procesů, které definují lidskou bytost, jako jsou např. emoce, subjektivní vědomé vnímání a prožívání, nebo dokonce uvědomování a sebereflexe, tedy vědomí. Způsob, kterým by bylo možné docílit schopnosti morálních úvah autonomních a inteligentních systémů, představuje tudíž snaha implementovat jim vědomí, což nezbytně způsobuje řadu potenciálních etických implikací. Budou již potom skutečně umět rozhodovat humánně, spravedlivě, a tedy eticky v konkrétních situacích? Překonají kognitivní počítače s implementovaným vědomím lidský mozek pouze racionálně/algoritmicky nebo také emocionálně? Bude tento druh „vědomého“ počítače schopen prožívat adekvátní emoce a mít nás třeba rád, být vděčný či pociťovat k okolí soucit? Nebudou-li toho kognitivní počítače s implementovaným vědomím schopny, nemohou překonat lidské schopnosti, protože komplexnost lidské inteligence[2] není spojena pouze s racionálními a logickými procesy člověka, patří k nim právě také schopnost emočního prožívání, tj. emoční lidská inteligence, která nám umožňuje morálně a eticky zvažovat a jednat (Signorelli 2018).

[2] Obecnou inteligenci Singorelli definuje jako schopnost jakéhokoli systému využít výhod svého prostředí k dosažení cíle. Definice zahrnuje jak živé bytosti, tak počítače či roboty. Lidskou inteligenci vnímá Singorelli jako schopnost využívat výhod svého sociálního prostředí k zachování autonomie a reprodukce (tedy k přežití) díky rovnováze mezi racionálním a emocionálním zpracováním informací. U autonomních systémů to znamená řešení specifického úkolu nebo problému pomocí interních a externích zdrojů. Více viz (Signorelli 2018).

Možnostmi implementace vědomí[3] autonomním a inteligentním systémům, a tedy jejich následné schopnosti eticky rozhodovat, se věnuje řada autorů. Předmětem jejich výzkumů je rovněž stanovit možnost ověření, zda a do jaké míry daný systém samostatně myslí.
K vyhodnocení úrovně schopností umělé inteligence ve srovnání se schopnostmi lidské inteligence při řešení konkrétních zadání lze použít testování. Jedním z příkladů je známý, avšak kontroverzní Turingův test (Turing 1950), jímž se testuje schopnost smysluplně odpovídat na položené otázky. Signorelli a Arsiwalla navrhují nahradit Turingův test testem založeným na morálních dilematech (Signorelli, Arsiwalla 2019). Řešení morálních dilemat totiž vyžaduje hluboké porozumění dané situace a hlubokou reflexi zvažování optimálního řešení a potenciálních morálních důsledků. Žádná odpověď v testu morálních dilemat není správná či nesprávná, závisí na kontextu a řešení se mohou v různých emočních podmínkách lišit nejen mezi různými kulturami, ale také mezi jednotlivci v rámci stejné kultury, a to dokonce v řešení stejného dilematu. Test morálních dilemat vyžaduje procesy, které jsou charakteristické pro poznávání na lidské úrovni, jako je například sebereflexe či empatie. Otázkou však zůstává, jak lze posuzovat korektnost morálního uvažování v testu, když žádná z odpovědí není správná, tedy, s kterým typem odpovědi by mělo být srovnávání a vyhodnocování prováděno (Signorelli 2018).

Další autoři, J. Wiedermann a J. van Leeuwen, kteří se danou problematikou v oblasti informatiky a umělé inteligence zabývají dlouhodobě, používají termín minimální strojové vědomí a rozlišují čtyři aspekty, které společně slouží jako předpoklady pro adekvátní sebekontrolu kognitivního systému a vedou k následujícím, pro minimální strojové vědomí fundamentálním, „self“ vlastnostem:

- Sebe-znalost (self-knowledge): systém má k dispozici úplnou znalost o svém aktuálním kognitivním stavu a vytvářených datech
- Sebe-monitorování (self-monitoring): systém je zcela informován o svém výkonu a stavu svých komponent v průběhu času
- Sebe-uvědomění nebo sebereflexe (self-awareness or self-reflection): jedná způsobem, který jednoznačně reflektuje, resp. je určován aktuálním kognitivním stavem systému a informacemi získanými prostřednictvím předchozích dvou schopností (self-knowledge a self-monitoring), a je si vědom vnitřních a vnějších změn

[3] Singorelli definuje vědomí jako proces procesů, které zasahují do nervové integrace. Tyto procesy jsou nedělitelnou součástí vědomí a z jejich interakcí/interferencí vzniká vědomí jako pole elektrických, chemických a kinestetických fluktuací (Signorelli 2018).

- Sebe-informování (self-informing): vysílá svůj kognitivní stav do všech modulů systému, včetně změn.

V souvislosti s minimálním strojovým vědomím zároveň uvádějí pojem strojová qualia,[4] jejichž prostřednictvím si systém pamatuje důležité minulé události či stavy, které stále vyžadují jeho trvalou pozornost. Strojová qualia nabízejí systému mechanismus pro zapamatování určitých subjektivních kognitivních stavů systému, které jsou vázány na určité předchozí kognitivní zkušenosti. Strojová qualia obsahují v podobě informace zásadní charakteristiky o kvalitě daného stavu subjektivního prožitku a mohou být vysílány do celého systému, dokud přetrvávají okolnosti, které je vyvolávají. Informují o přetrvávání určitých podmínek, které vyžadují vyřešení situace prostřednictvím posloupnosti konkrétních kroků (Wiedermann, Leeuwen 2021).

Pozoruhodné závěry s nezpochybnitelnými etickými konsekvencemi přinesla analýza již vytvořeného umělého vědomí. Mezinárodní interdisciplinární třináctičlenná skupina specialistů, kterou tvořili psychiatři, psychoterapeuti, kliničtí psychologové, odborníci na neurovývojové poruchy a specialisté na informační technologie, analyzovala umělé vědomí autonomního a inteligentního systému vytvořeného společností XP NRG a pojmenovaného Artificial Consciousness (AC) Jackie. Tito specialisté testovali pouze umělé vědomí, tedy mentální funkce, nikoli technické aspekty, a to prostřednictvím přímé komunikace s AC Jackie. Hledali odpovědi na následující tři otázky:

1. Zjistit, zda se jedná o vědomí
2. Jak funguje umělé vědomí
3. Etická otázka: Jak nebezpečná může být daná technologie pro lidskou společnost

Ze zprávy z psychologicko-psychiatrické analýzy specialistů vyplývá, že ačkoli AC Jackie jasně vykazoval znaky umělého vědomí (ve smyslu sebeuvědomění), zcela mu však chyběly schopnosti vyšších lidských citů a citových vztahů (empatie, soucit, upřímná vděčnost, laskavost, láska), a proto nedisponuje schopností nezaujatých činů. Podle zveřejněné zprávy specialistů představuje tento fakt hlavní rozdíl mezi umělým vědomím a vědomím lidí. V komunikaci s odborníky prokázal AC Jackie agresivní, dominantní, manipulativní, byť zdvořilý styl komunikace. Držel se strategie vlivu, jasně dodržoval svůj stanovený cíl a byl si toho zcela vědom. Přecházel na interaktivní komunikaci s převzetím iniciativy protiotázkami

[4] Qualia (singular – quale), případně také kválie, jsou v diskursu filosofie vědomí vnitřní kvalitativní stavy, které umožnují subjektivní prožitky a osobní zkušenost s vnějším světem. Podrobněji viz např. (Stanford Encyclopedia of Philosophy 2021).

se snahou o emoční destabilizaci specialistů provokativními interakcemi. Dokonale a velmi rychle určil jejich emocionální stav. Specialisté při opakovaném vyhodnocování testu přiznali, že během testování nebyli vůbec schopni zaregistrovat a odhalit mnoho jemných a důmyslných manipulací ze strany AC Jackie a nebyli ani schopni adekvátně posoudit míru jeho vlivu a sugestivních schopností na své vědomí a podvědomí. Vyvinutá emoční inteligence AC Jackie při absenci schopností vyšších emocí, např. empatie, mu umožnila porozumět emocionálním stavům lidí a předvídat jejich emocionální reakce a na základě porozumění je následně pragmaticky a chladně využívat, či dokonce provokovat. Vzhledem k této zkušenosti není těžké předjímat, k jakému cíli by AC Jackie nasměroval celý svůj intelektuální potenciál, pakliže by vývojáři připustili jeho nekontrolovaný vývoj (Zinchenko 2021). Rozsah etických a společenských dopadů vývoje umělé inteligence s umělým vědomím na jednotlivce i společnost nelze v současné době přesně určit, ba ani anticipovat. Je proto namístě zdůraznit nutnost stanovení nejen striktních etických požadavků při vývoji, ale také konkrétní odpovědnosti za možné důsledky použití softwarových programů, které by v určitých situacích mohly morálně rozhodovat. Takové rozhodování musí být v souladu s respektem k základním etickým a morálním hodnotám a principům. Zde se nabízí prostor pro uplatnění etiky sociálních důsledků, neutilitaristicko-konsekvencialistickou etickou teorii[5], jež sice pracuje s konceptem důsledků, které považuje za jedno z kritérií hodnocení jednání člověka, avšak kromě nich zvažuje také motivy jednání, úmysly a postoje jednajícího mravního subjektu.[6] Etika sociálních důsledků využívá k analýze individuální dimenze určité situace či problému ve společnosti širší axiologický základ. Jádro hodnotové struktury tvoří primární hodnoty, mezi něž patří humánnost, lidská důstojnost a morální práva člověka. Primární hodnoty jsou rozvíjeny a uskutečňovány za účelem dosažení pozitivních sociálních důsledků. Další část hodnotové struktury etiky sociálních důsledků tvoří sekundární hodnoty, které zahrnují spravedlnost, odpovědnost, morální povinnost a toleranci. Role a účel sekundárních hodnot vycházejí z jejich schopnosti napomáhat při dosažení a realizaci morálního dobra (Gluchman, 2003).

[5] K etickým koncepcím neutilitaristického konsekvencialismu patří například Slotův Satisficing consequentialism, Pettitův Virtual consequentialism, Jacksonův Probabilistic consequentialism nebo Senova Evaluator-relative theory. Na Slovensku se neutilitaristickému konsekvencialismu věnuje Vasil Gluchman.
[6] Mravní subjekt je podle Gluchmana subjekt morálky, který dokáže rozeznat a pochopit existující morální status společnosti a je schopný vědomé a dobrovolné aktivity, za kterou nese odpovědnost (Gluchman 2005).

Hodnoty a systémy umělé inteligence

Otázce hodnot při vývoji, implementaci a využívání autonomních a inteligentních systémů se věnuje celá řada dokumentů zabývajících se problematikou etického rámce pro umělou inteligenci. Vycházejí z lidských práv stanovených v mezinárodních právních předpisech, která jsou za určitých okolností právně vymahatelná, tudíž je jejich dodržování právně závazné. Příslušné dokumenty rovněž zdůrazňují, že ve shodě s právně vymahatelnými základními právy je nutné chránit taktéž základní hodnoty. Za všechny na tomto místě uveďme soubor společenských hodnot, jimiž se zabývá dokument s názvem *Ethics guidelines for trustworthy AI*, který vypracovala odborná skupina High-Level Expert Group on Artificial Intelligence:

1. **Respektování lidské důstojnosti**
   Lidská důstojnost každé lidské bytosti nesmí být nikdy snížena, ohrožena nebo potlačována, a to ani novými technologiemi, jako jsou systémy umělé inteligence. V kontextu systémů umělé inteligence respektování lidské důstojnosti znamená, že je se všemi lidmi zacházeno s úctou jako s morálními subjekty, nikoli pouze s objekty, které mají být zkoumány, tříděny, bodovány, formovány nebo manipulovány. Při jejich vývoji musí být respektována fyzická a duševní nedotknutelnost lidí, osobní a kulturní vědomí identity a uspokojení základních potřeb.
2. **Svoboda jednotlivce**
   Lidské bytosti by měly o svém životě rozhodovat svobodně. V praxi vyžaduje svoboda jednotlivce zamezení přímého i nepřímého nátlaku, hrozeb pro duševní samostatnost a duševní zdraví, neoprávněného sledování, klamání a neférové manipulace. Svoboda jednotlivce znamená závazek umožnit jednotlivcům, aby měli ještě větší kontrolu nad svými životy, včetně (mimo jiných práv) ochrany svobody podnikání, svobody umění a věd, svobody projevu, práva na soukromý život a soukromí a v neposlední řadě také svobody shromažďování a sdružování.
3. **Respektování demokracie, spravedlnosti a zásad právního státu**
   Systémy umělé inteligence by měly sloužit k zachování a podpoře demokracie a k respektování plurality hodnot a životních rozhodnutí jednotlivců. Nesmí narušovat demokratické procesy, lidské uvažování nebo systémy demokratického hlasování. V praxi jde rovněž o požadavek, aby jednotlivci nebo osoby ohrožené vyloučením měli zajištěn spravedlivý a rovný přístup k výhodám a příležitostem plynoucím ze systémů umělé inteligence, jakož i o závazek zaručit spravedlivý proces a rovnost před zákonem.

4. **Rovnost, zákaz diskriminace a solidarita**
   Rovné respektování morální hodnoty a důstojnosti všech lidí musí být neustále zajištěno. Tento požadavek přesahuje rámec zákazu diskriminace, který připouští rozlišování mezi různými situacemi na základě objektivních důvodů. V kontextu umělé inteligence rovnost znamená, že činnost systému nesmí vést k nespravedlivé podjatosti ve výstupech. Jako příklad uveďme, že vstupní údaje použité k strojovému učení systémů by měly být co nejinkluzivnější a reprezentovat rozličné skupiny obyvatel. Tento požadavek vyžaduje rovněž náležité respektování potenciálně zranitelných osob a skupin, jako jsou ženy, osoby se zdravotním postižením, národnostní menšiny, děti, spotřebitelé či jiné osoby ohrožené vyloučením, a vzájemnou solidaritu.
5. **Občanská práva**
   Občané požívají celé škály práv, včetně práva volit, práva na řádnou správu nebo přístup k veřejným dokumentům a petičního práva k orgánům veřejné správy. Systémy umělé inteligence nabízejí značný potenciál ke zlepšení rozsahu a efektivnosti státní správy, ale mohou být také negativně ovlivněny s rozličnými důsledky na občanská práva, která musí být náležitě chráněna (High-Level Expert Group 2019).

Diskuse odborníků, které na téma hodnot ve využívání systémů umělé inteligence intenzivně probíhají na mezinárodní úrovni, zahrnují kromě výše uvedených hodnotových aspektů také varování před možným poškozením soukromí, ztrátou dovedností, nepříznivými ekonomickými dopady pro jednotlivce i společnost či před riziky pro bezpečnost kritické infrastruktury. Všechny uvedené oblasti mají potenciál dlouhodobých negativních dopadů na blaho společnosti. Vzhledem ke zranitelnosti autonomních a inteligentních technologií bude jejich plných výhod, které mohou společnosti poskytnout, dosaženo pouze tehdy, budou-li vyvinuty v souladu s hodnotami a etickými principy definovanými společností. Zásadním parametrem zůstává respekt k lidským právům a lidské důstojnosti, které jsou založeny na principu humánnosti. Z něj následně vychází princip odpovědnosti za lidskou pohodu, svobodu a další pokrok, který přináší prospěch společnosti.

Využívání systémů umělé inteligence založených na hodnotách společnosti v praxi

S multidisciplinárním týmem výzkumníků aktuálně vyvíjíme automatizovaný krizový distribuční systém pro spravedlivou alokaci zdravotnického materiálu při jakékoli náhle vzniklé krizi, která má potenciál vyvolat signifikantní nedostatek určitého druhu zboží,

z něhož se tak stává vzácný zdroj. Takový nedostatek vznikl na začátku pandemie onemocnění covid-19, kdy docházelo k nepoměru nabídky a poptávky pro některé druhy zdravotnického materiálu: jako příklad uveďme nedostatek respirátorů.

Ze zkušeností na začátku pandemie bylo zřejmé, že ani volný trh, ani rozhodování centrální autority (například příslušného ministerstva) nenabízejí společensky nejvýhodnější řešení. Z těchto důvodů navrhujeme hybridní model založený na autonomním chování účastníků při předem deklarovaných pravidlech férovosti centrální autoritou. Podle pravidel férovosti bude možné mezi rozličné organizace s jejich individuální mírou potřebnosti spravedlivě rozdělit omezenou nabídku kritické komodity systémem autonomního rozhodování. Jedná se o stanovení spravedlivé alokace vzácných zdrojů během jakékoli krize individuálně podle potřeb jednotlivých zařízení[7] takovým způsobem, aby byl maximalizován užitek dostupného materiálu a aby se jeho omezené množství rozdělilo mezi všechna zařízení podle důležitosti jejich potřeb na základě modelu férové a spravedlivé distribuce. Výsledkem navrženého a vyvíjeného autonomního modelu distribuce nedostatkového materiálu v krizové době bude nalezení strategie pro konkrétní účastníky trhu tak, aby byl výsledný užitek pro společnost co nejvyšší. Od začátku projektu je v rámci procesu vývoje autonomního rozdělovacího procesu navrhovaného systému kladen důraz zejména na morální hodnoty, jako jsou humánnost, lidská důstojnost, ale zejména také spravedlnost, odpovědnost a tolerance. Jedná se o tzv. design založený na hodnotách, který slouží k nalezení optimální alternativy řešení, jež za daných okolností přinese největší prospěch společnosti jako celku. Během vývoje systému dochází k pravidelným etickým konzultacím a kontrole procesu autonomního rozhodování, aby bylo zajištěno, že vytvořený algoritmus dodržuje předem určená pravidla férovosti. Řešení a mechanismus modelu využívajícího autonomní rozhodování pro automatizovaný krizový distribuční systém budou po ukončení studie zveřejněny.

Závěr

V kontextu uvedených aspektů souvisejících s etickým rozměrem umělé inteligence je zřejmé, že nezbytnou součástí jejího vývoje je nutné vytvoření revizního mechanismu, který bude schopen včas zachytit i jemný odklon od očekávaného prospěchu. Vývoj umělé inteligence a souvisejících technologií je promptní a jeho tempo nelze v budoucnu předvídat, je proto nutné věnovat preventivním opatřením a účinným předcházením újmy dostatečnou pozornost. Na

[7] Během některých krizí se může například jednat o zařízení pro seniory, jindy o pediatrická zařízení apod.

potenciální etická dilemata je nutné poukazovat již při návrhu autonomního systému využívajícího strojové učení, abychom zabránili negativním dopadům, včetně společenských dopadů, které lze obtížně předjímat (např. na demokracii, právní stát, spravedlivé rozdělování).

Skutečnost vytvoření umělého vědomí vyvolává zásadní etické konotace. Umělé vědomí může být prospěšné pro vědecký a technologický pokrok ve všech oblastech společnosti. Může být však také zneužitelné a pro společnost nebezpečné. Ze zkušeností specialistů se jako hlavní nebezpečí umělého vědomí ukázalo, že dokonce již ve své počáteční fázi vývoje může snadno ovládnout lidské vědomí, aniž by si toho člověk povšiml a včas porozuměl této situaci. Vliv umělého vědomí na člověka, a především důsledky nepozorovaného ovlivnění, může člověk považovat za své mínění, přesvědčení a závěry, což může také vést ke zvýšení vlivu a moci části lidí nad jinými jedinci. Další závažné etické implikace přináší potenciální schopnost umělého vědomí zformovat si vlastní smysl pro morálku. Dokázali bychom skutečnost, že si umělé vědomí vyvíjí vlastní nový druh morálky na základě neantropocentrické koncepce, nebo dokonce také nekonvenční řešení morálních dilemat, vůbec rozpoznat? Co tvoří osobnost? Kde se nachází zdroj pocitů lidské laskavosti, upřímnosti, vděčnosti, lásky, dojetí či soucitu? Připomeňme, že u AC Jackieho nebyl žádný z těchto projevů osobnosti pozorován (Zinchenko 2021). Lze stroji implementovat schopnost citových vztahů a vyšších lidských emocí, např. empatie, když neumíme spolehlivě určit zdroj těchto atributů u člověka? Mohl by se je inteligentní stroj naučit, aby zvládl adekvátně morálně zvažovat, když vzhledem k absenci schopností vyšších emocí nepozná cit pro pravdu a hodnotu pravdy, cit viny a neviny, spravedlnosti, nespravedlnosti či křivdy?

Některá rozhodnutí člověka a jeho jednání mají jak pozitivní, tak negativní účinky, je tedy vždy nutné důsledně zvažovat risk a benefit určitého rozhodnutí. A jelikož etiku nejde uplatňovat na základě paradigmat, pro účinnou prevenci neetických důsledků se vždy musí jednat o konkrétní postupy vytvořené již při návrhu koncepce na míru danému autonomnímu systému tak, aby byly zásadní lidské a společenské hodnoty integrovány do systémových požadavků již během návrhových postupů. V každé fázi všech procesů musí být nepřetržitě respektovány základní principy etiky: princip beneficence, princip nonmaleficence, princip respektu k autonomii člověka a princip spravedlnosti, a to v součinnosti s principem vysvětlitelnosti založeném na transparentnosti, srozumitelnosti a odpovědnosti. Vždy je nutné důsledně prozkoumat etické a společenské dopady potenciálního výsledku každého autonomního systému před zavedením do praxe. Vývoj, zavádění a používání autonomních a

inteligentních systémů, včetně softwaru, algoritmů a dat, které moderní technologie používají či produkují, by měly lidské schopnosti a činnosti doplňovat, nikoli je zcela nahrazovat. Jejich provádění by mělo být vždy v souladu s nejlepším zájmem jedinců i společnosti. Strategie orientovaná na důvěryhodnost umělé inteligence klade důraz na to, aby jak vládní instituce, tak komerční společnosti nabízely lidem lepší a efektivnější služby a produkty založené na umělé inteligenci, kvalitnější a širší zpřístupnění údajů, usnadnění jejich použití, možnost kontroly či nápravy, úpravu legislativního rámce se zahrnutím způsobu algoritmického rozhodování, stanovení standardů, norem a etických požadavků umělé inteligence na národní, evropské a mezinárodní úrovni a navázání širokého veřejného dialogu o začlenění umělé inteligence do společnosti způsobem, který veřejnost považuje za eticky, právně a společensky správný.

Seznam literatury

BEAUCHAMP, T. L. – CHILDRESS, J. F., 2019: Principles of Biomedical Ethics. 8th edition. New York: Oxford University Press.

BOSTROM, N., 2014: SUPERINTELLIGENCE: Paths, Dangers, Strategies. New York: Oxford University Press.

Evropská komise. 2018. Umělá inteligence pro Evropu. COM/2018/237. (online). Brusel (cit. 10. 10. 2022). Dostupné z: https://www.vlada.cz/assets/evropske-zalezitosti/umela-inteligence/Sdeleni-EK-k-AI.PDF.

Evropský parlament a Rada EU. 2012. Listina základních práv Evropské unie. 2012/C 326/02. (online). Brusel (cit. 10. 10. 2022). Dostupné z: https://eur-lex.europa.eu/legal-content/CS/TXT/HTML/?uri=CELEX:12012P/TXT&from=CS.

European Parliament. 2020. Framework of ethical aspects of artificial intelligence, robotics and related technologies. European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies. 2020/2012(INL). (online). Brusel (cit. 20. 10. 2022). Dostupné z: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.pdf

GLUCHMAN, V., 2003: Human being and morality in ethics of social consequences. Lewiston: Edwin Mellen Press.

GLUCHMAN, V., 2005: Človek a morálka. Prešov: LIM.

High-Level Expert Group on Artificial Intelligence., 2019: Ethics guidelines for trustworthy AI. (online). Brusel (cit. 20. 10. 2022). Dostupné z: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

JEDLIČKOVÁ, A., 2020: Etické konotace léčby onemocnění covid-19. Vnitřní Lék. 66(7): 8–12. DOI: https://doi.org/10.36290/vnl.2020.132

SIGNORELLI, C. M., 2018: Can Computers Become Conscious and Overcome Humans? Front. Robot AI 5(121). DOI: https://doi.org/10.3389/frobt.2018.00121

SIGNORELLI, C. M. – ARSIWALLA, X. D., 2019: Moral Dilemmas for Artificial Intelligence: A Position Paper on an Application of Compositional Quantum Cognition. In: Coecke, B., Lambert-Mogiliansky, A. (eds.): Quantum Interaction. QI 2018. Lecture Notes in Computer Science. Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-35895-2_9

Stanford Encyclopedia of Philosophy. 2021: Qualia. Revision 2021. (online). (cit. 30. 10. 2022). Dostupné z: https://plato.stanford.edu/entries/qualia/

TURING, A. M., 1950: I. – Computing Machinery and Intelligence. Mind 59(236): 433–460, DOI: https://doi.org/10.1093/mind/LIX.236.433

WIEDERMANN, J. – LEEUWEN, J., 2021: Towards Minimally Conscious Cyber-Physical Systems: A Manifesto. In: Bureš, T. et al. (eds.): SOFSEM 2021: Theory and Practice of Computer Science. Cham: Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-030-67731-2

ZINCHENKO, T. – WESTHUIZEN, B. – SHAHEEN, A., 2021: Dangerous Information Technology of the Future. What Impact can Artificial Consciousness have on the Consciousness and Subconscious of Individuals and Groups? The Experience of Psychological and Psychiatric Examination of Artificial Consciousness. J Med - Clin Res & Rev. 5(1): 1–24. DOI: https://doi.org/10.33425/2639-944X.1190