ETICKÁ PRAVIDLA REGULACE AI

Centrum Karla Čapka pro studium hodnot ve vědě a technice

OBSAH

SITUACE V ČR A VE SVĚTĚ

ÚVOD

Umělá inteligence (dále jen AI) ve svých různých podobách, ztělesněních (roboti) a aplikacích slibuje výrazné benefity pro lidské jedince uvažované jednotlivě a jako společnost. Má výrazný potenciál přispět k teoretickému a aplikovanému výzkumu, zlepšit distribuci zdrojů, podpořit růst průmyslu a zlepšit jeho efektivitu, stále větší úlohu bude hrát v každodenním lidském životě v podobě medicínských robotů, medicínských expertních a diagnostických systémů, jako řídící algoritmy autonomních vozidel, průmyslových robotů, domácích, asistenčních či záchranných robotů.

Klíčovým momentem rozvoje a využívání systémů umělé inteligence ve společenském měřítku je požadavek, aby nedocházelo k narušení struktury společnosti a složitého přediva vazeb, díky nimž je společenství lidí skutečnost lidskou společností. Tento požadavek je možné vyjádřit pojmem **důvěryhodné umělé inteligence** (*Trustworthy AI*), která splňuje dva klíčové požadavky:

1. Ve svém vývoji a využití respektuje základní lidská práva, hodnoty a etické principy a podléhá systému regulace, jenž zajišťuje, že se AI chová eticky korektním způsobem.

2. Je z technického hlediska dostatečně robustní a spolehlivá.

Respekt k těmto dvěma principům zajišťuje, že stále širší využívání systémů umělé inteligence v praxi nepovede k morální panice, vzrůstu obavy z AI a odmítání jejích aplikací ve společnosti. Je tedy zřejmé, že etická a hodnotová dimenze výzkumu, vývoje a aplikace AI není něčím, co by vstupovalo do hry v pozdějších fázích tohoto vývoje; představuje hodnotový rámec, který musí být inkorporovaný ve všech jeho okamžicích, od výzkumu, návrhu designu a algoritmů, až po jejich implementaci a praktické využití.

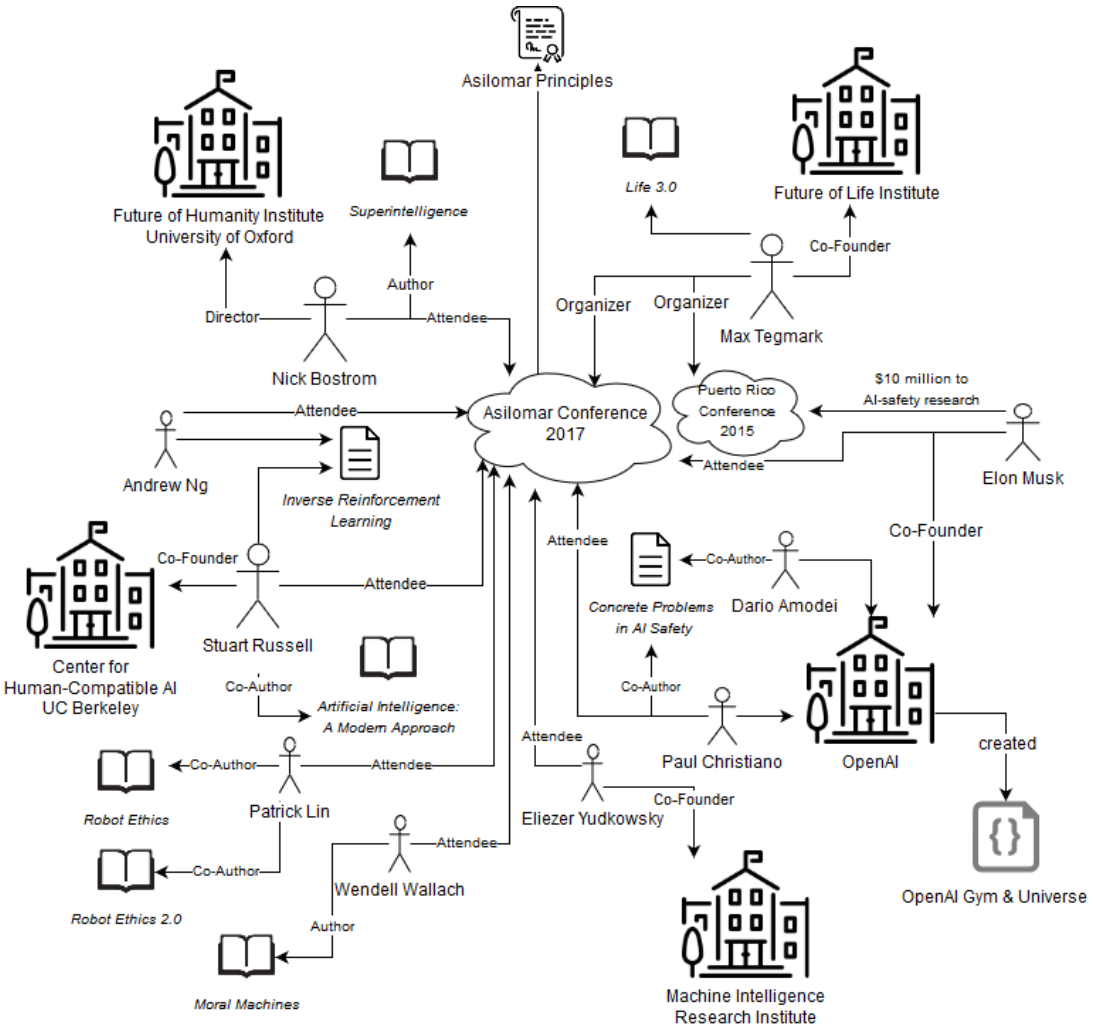
SITUACE VE SVĚTĚ

Zvláště v zemích EU a USA je hodnotové a etické dimenzi vývoje a využití AI již delší dobu věnována velká pozornost, což se v praxi projevuje několika způsoby.

1. Existuje celá řada výrazných odborníků věnujících se etice AI, roboetice a filosofii techniky. Mezi nejdůležitější postavy rozvíjející hodnotovou a etickou dimenzi AI patří Nick Bostrom, Stuart Russell, Wendell Wallach, Patrick Lin, Dario Amodel či Eliezer Yudkowsky. Výsledkem práce těchto odborníků jsou publikace pokládající základy etiky AI, např. *Superintelligence*, *Robot Ethics*, *Robot Ethics 2.0*, *Moral Machines*, *Life 3.0* či *Towards a Code of Ethics for Artificial Intelligence*. Všichni zmínění odborníci se rovněž v roce 2017 zúčastnili klíčové konference Asilomar Conference, která předložila celou řadu důležitých etických principů regulace výzkumu a využívání AI.
2. V poslední době vznikla také celá řada multidisciplinárních center zabývajících se výzkumem etiky AI. Mezi nejvýznamnější patří Future of Humanity Institute, Future of Life Institute, Center for Human-Compatible AI, OpenAI, Machine Intelligence Research Institute, Leverhulme Centre for the Future of Intelligence či centrum Philosophy of Media and Technology na Vídeňské univerzitě.

3. Na zahraničních univerzitách existují programy Ph.D. studia v oblasti aplikované etiky, filosofie techniky a AI a etiky umělé inteligence. Tyto programy umožňují interdisciplinární výzkum hodnotové a etické dimenze moderních technologií a vychovávají odborníky, kteří se mohou podílet na výzkumu a vývoji věrohodné AI a spolupracovat s ostatními odborníky na vývoji etických směrnic regulujících tento výzkum a jeho aplikace v praxi.

Situaci ve světě v oblasti etiky AI názorně shrnuje následující obrázek:



18. 12. 2018 publikovala High-Level Expert Group on Artificial Intelligence při Evropské komisi návrh etických směrnic pro věrohodnou AI (*Ethics Guidelines for Trustworthy AI*), které ukazují, že v rámci EU je tomuto problému věnována velká pozornost. Tento návrh by se

v ČR měl stát východiskem k širší odborné a společenské diskuzi uzavřené vytvořením a publikací směrnic platných v ČR.

SITUACE V ČR

V České republice je v porovnání se zahraničím výrazně horší. Existuje zde pouze jeden doktorský program studia aplikované etiky (FHS UK, prof. Haškovcová), zaměřuje se však téměř výhradně na lékařskou etiku. Rovněž v doktorských programech bioetiky (1. LF UK, LF MU Brno) se vzdělávají především odborníci na bioetiku a lékařskou etiku. Důsledkem tohoto stavu je to, že v České republice je jen velmi málo odborníků, kteří se mohou odborně věnovat problematice etiky AI.

V nedávné době neexistovalo žádné interdisciplinární centrum zaměřené na hodnotou a etickou dimenzi AI a moderních technologií. Tato situace se změnila na podzim roku 2018, kdy při slavnostním podpisu smlouvy o spolupráci mezi Ústavem informatiky AV ČR, Filosofickým ústavem AV ČR, Ústavem státu a práva AV ČR a Přírodovědeckou fakultou UK vzniklo meziústavní Centrum Karla Čapka pro studium hodnot ve vědě a technice. Třebaže toto centrum existuje velmi krátkou dobu, podařilo se mu již navázat rozsáhlou mezinárodní spolupráci (mezi členy centra patří např. prof. Patrick Lin), díky níž bude centrum v červnu 2019 pořádat mezinárodní konference o etice autonomních vozidel. Členové centra také připravují publikace o etice autonomních vozidel, která se připravuje se spolupráci s nakladatelstvím Oxford University Press.

Rovněž v oblasti tvorby etických směrnic regulujících výzkum, vývoj a aplikace AI Česká republika výrazně pokulhává za EU a USA. Žádné směrnice neexistují ani v podobě prvotního návrhu určeného k širší odborné diskusi, nebyla ustavena žádná odborná komise, která by se tomuto problému věnovala a splňovala by podmínku odbornosti svých členů, tj. schopnosti orientovat se v současné etice a moderních technologií.

ZÁKLADNÍ ETICKÉ SYSTÉMY

V současné normativní etice existuje celá řada etických systémů, které se teoreticky nabízejí jako nástroje regulace AI. Jako nejvhodnější lze určit následující dva:

- Konsekvencialistické přístupy
- Deontologické přístupy

KONSEKVENCIALISTICKÉ PŘÍSTUPY

Konsekvencialistické etické teorie vycházejí z předpokladu, že jediným morálně relevantním normativním faktorem – tj. faktorem určujícím, zda určité jednání je z morálního hlediska správné či nesprávné – jsou důsledky jednání. V případě AI lze hovořit o „umělé morálce", tj. o hodnocení důsledků jednání AI a strojů, které disponují umělou inteligencí, aniž bychom museli brát v úvahu další důležité problémy, jakým je např. morální odpovědnost.

Je-li *X* systém umělé inteligence schopný jednání, *J* představuje určité jednání/činnost/aktivitu, a jsou-li $D_1, \ldots, D_n$ důsledky tohoto jednání, potom o morálním hodnocení (správné-nesprávné) *J* rozhodují pouze důsledky $D_1, \ldots, D_n$.

Příklad: Autonomní vozidlo řídící se konsekvencialistickou etikou minimalizace újmy v situaci možné kolize zváží: i) možné modality jednání (zahnout vlevo, vpravo, pokračovat v jízdě, prudce brzdit a neměnit směr…), ii) předpokládané důsledky těchto modalit jednání, iii) předpokládanou celkovou újmu spojenou s těmito důsledky, iv) určení modality jednání, která přináší nejnižší újmu, v) provedení tohoto jednání. Nakonec provede akci, jejíž důsledky přinesou nejnižší možnou újmu všem zúčastněným stranám.

Etika založená na zvažování důsledků musí disponovat klíčem, podle něhož budeme důsledky zvažovat, tj. musí disponovat možností hodnotového (dobré-lepší-špatné-horší) posouzení důsledků jednání.

Příklad: Etické algoritmy autonomních vozidel musí zahrnovat tři kategorie možné újmy (na majetku – na zdraví – na životech) a způsoby, jak tuto újmu kvantifikovat a srovnávat nejen v rámci jedné kategorie, ale i napříč kategoriemi.

Přitažlivost konsekvencialistické etiky, zvláště v kontextu AI, spočívá v tom, že se nejedná o nějaký systém obecných morálních norem, které by AI musela interpretovat a rozhodovat v případě, že je na daný konkrétní případ možné aplikovat více protiřečících si norem.

Příklad. Představme si medicínského robota, který se řídí morálními normami „Minimalizuj utrpení pacienta", „Respektuj autonomní přání pacienta" a „Nesmíš usmrtit". V případě netěšitelné bolesti, kdy pacient autonomním způsobem opakovaně žádá o svou smrt, ocitají se normy „Respektuj autonomní přání pacienta" a „Nesmíš usmrtit" v konfliktu.

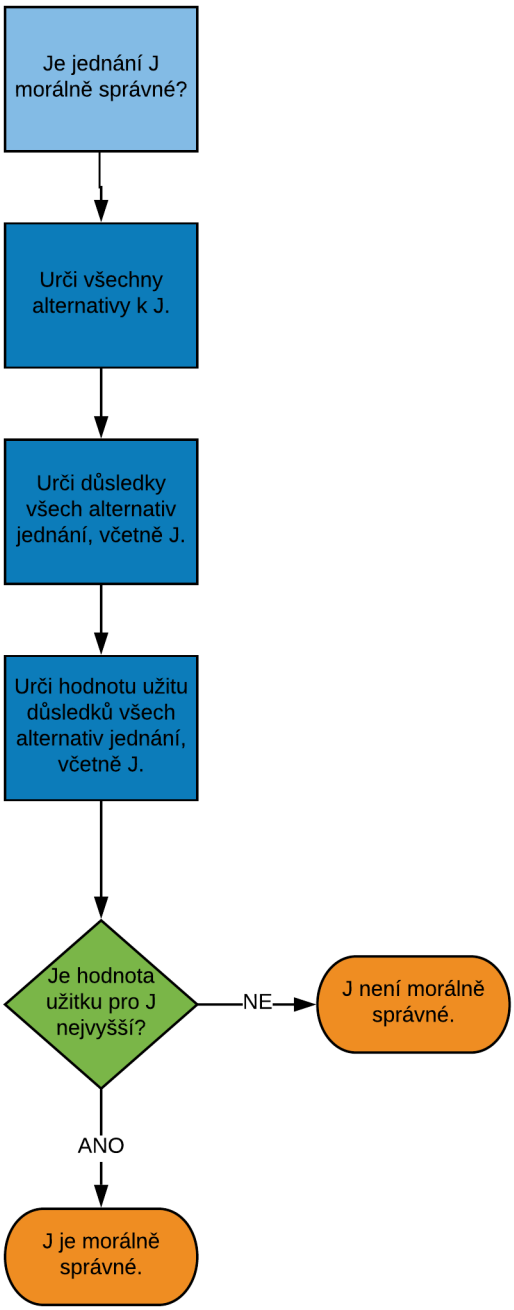
Jediným pravidlem – lze ho nazývat procesním pravidlem – konsekvencialistické etiky je pravidlo maximalizace užitku, jemuž můžeme dát následující podobu:

Je-li *X* systém umělé inteligence schopný jednání, J_1, … J_n představují možná jednání/činnosti/aktivity, potom morálně správné (a tudíž povinné) je jednání J_i, jehož důsledky přinášejí nejvyšší užitek.

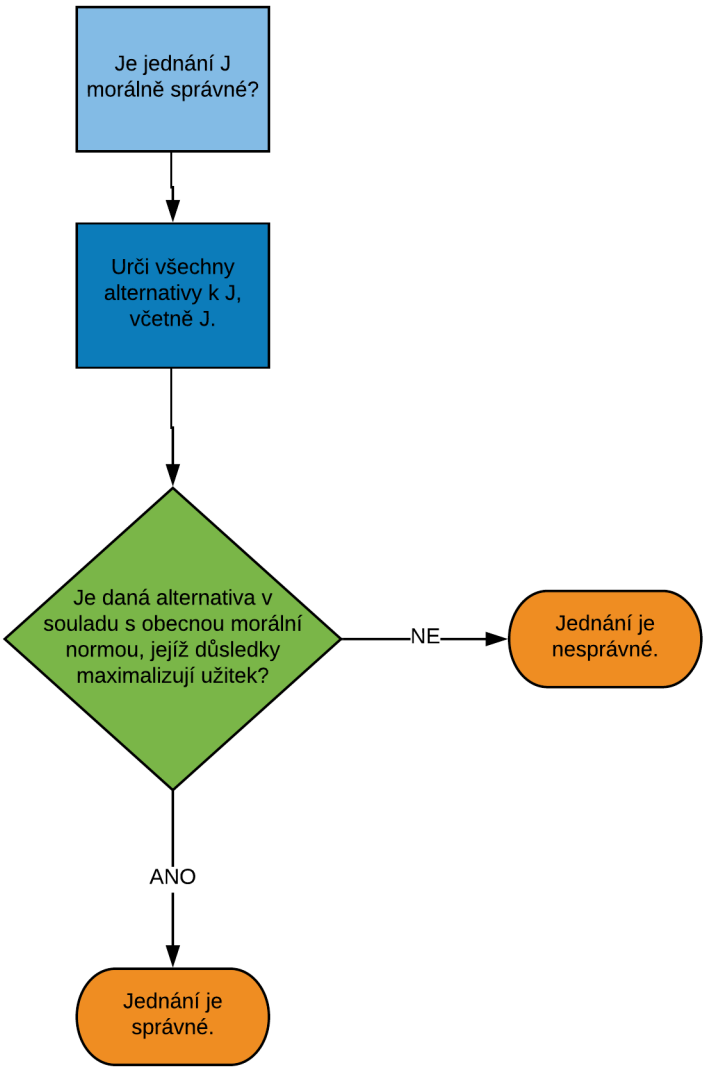
Určení parametrů hodnotového posouzení důsledků jednání AI (užitek) může být problematické zvláště v případě, kdy nehodnotíme negativní dopady její činnosti, nýbrž dopady pozitivní. Jakým způsobem maximalizovat užitek všech zúčastněných stran? Maximalizovat jejich štěstí, preference, potěšení? Neexistuje-li shoda na interpretaci pojmu „užitek", nelze konsekvencialistickou etiku uplatnit v praxi. Dalším problémem, ryze praktické povahy, je přesná kvantifikace užitku – jakým způsobem má AI určit, že její jednání způsobí určitý nárůst štěstí či potěšení?

Problém praktické neshody na interpretaci pojmu „užitek" a obtíže spojené s kvantifikací důsledků jednání ve shodě s nějakým hodnotovým klíček (které lidské bytosti provádějí intuitivně) představují hlavní překážky uplatnění konsekvencialistické etiky v regulaci AI.

Konkrétní rozhodovací proceduru v rámci konsekvencialistické etiky znázorňuje následující obrázek:



Konsekvencialistická praktická rozvahu může mít ještě jednu podobu. Namísto hodnocení důsledků každého konkrétního činu hodnotíme obecné morální normy a volíme takové, jejichž internalizace ve společnosti má nejlepší důsledky. Jednotlivé činy potom hodnotíme podle toho, zda spadají či nespadají pod zvolené morální normy. Tento proces shrnuje následující obrázek:

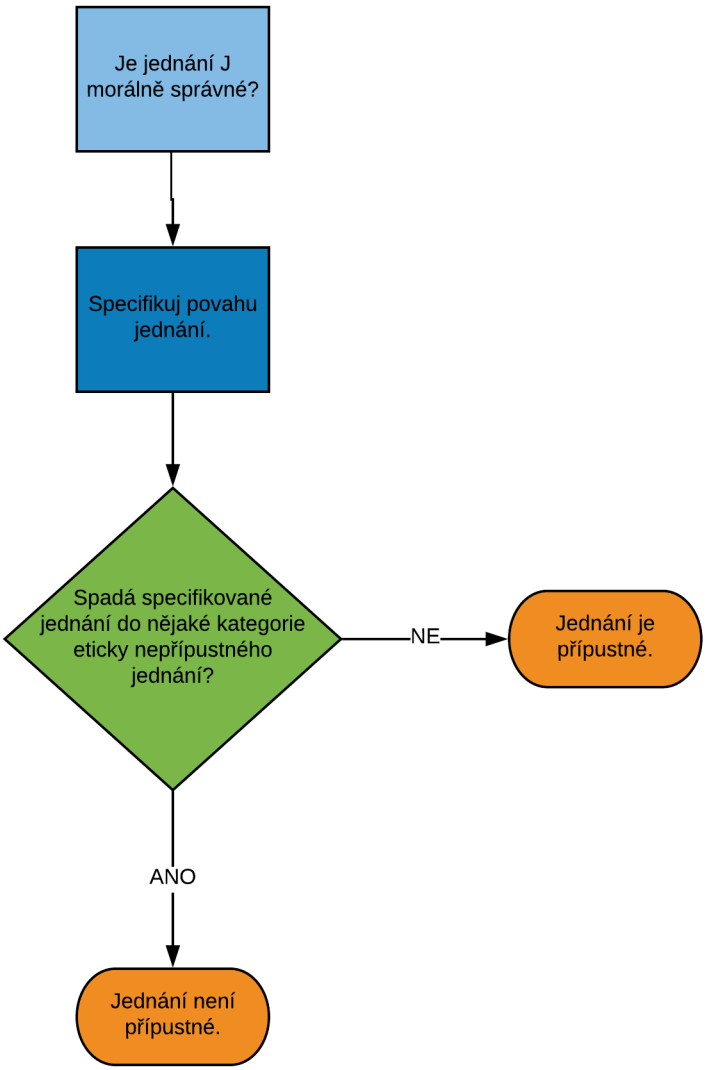


Vzhledem k obtížím, spojeným zvláště s interpretací užitku a kvantifikací důsledků jednání/činnosti AI, nepředstavuje konsekvencialistická etika vhodný etický systém pro regulaci AI obecně. Přesto však může existovat využití této etiky v rámci některých specifických aplikací AI, u nichž požadujeme jen omezený počet rozhodovacích alternativ a užitek je možné určit např. požadavkem na minimalizaci újmy. Tento etický systém – zřejmě v rámci určitých omezení – velmi pravděpodobně nalezne své uplatnění např. v řídících algoritmech autonomních vozidel.

DEONTOLOGICKÉ SYSTÉMY

Pod hlavičkou „deontologické systémy" se skrývá celá řada etických přístupů, které je obtížné jednoduše a jednoznačně charakterizovat. Obvykle pro ně platí, že pracují s obecnými etickými principy (např. respektem k lidské autonomii), které ukotvují celou řadu obecných morálních norem. Deontologické systémy se často používají při formulaci etických profesních norem, např. norem lékařské etiky, etiky soudců, novinářské etiky, roboetiky apod.

V případě deontologických etik se setkáváme s odmítnutím principu maximalizace užitku. Namísto něj se většinou postuluje celá řada obecných negativních a pozitivních morálních norem, které ukládají negativní (např. nezabiješ) či pozitivní (např. pomáhej lidem v nouzi) povinnosti. Rozhodovací procedura specifikující morální hodnocení určitého jednání je proto velmi odlišná od konsekvencialistických etik a může mít např. následující podobu:



Příklad: Medicínský robot zvažuje, zda odebrat orgány zdravému člověku a zachránit tak životy pěti nemocným a umírajícím pacientům. Takové jednání by sice maximalizovalo užitek, tento robot však postupuje podle pravidel deontologické etiky: usmrcení nevinné lidské bytosti spadá do kategorie „vražda" a vražda je typem jednání zapovězeném negativní morální normou „vražda je za všech okolností nepřípustná".

Deontologické přístupy k regulaci AI jsou více v souladu s mezinárodními úmluvami ukotvenými lidskými právy a respektem k lidské důstojnosti, svobodě, rovnosti a solidárnosti než etiky založené na konsekvencialistické rozvaze.

Pokud však chceme formulovat konkrétní obecné morální normy, musíme se shodnout na základních principech, o něž by se tyto normy

opíraly. V současnosti se v oblasti etiky AI nejčastěji zmiňují následující principy:

Princip beneficence
- Čiň dobro

Princip non-maleficence
- Nepůsob újmu

Princip autonomie
- Respektuj autonomii lidských bytostí a jejich aktérství

Princip férovosti
- Buď férový

Princip transparentnosti
- Jednej transparentně

Princip beneficence se pohybuje v rovině pozitivních morálních norem a závazků a požaduje, aby se AI chovala k lidem způsobem, který aktivně přispívá k rozvoji jejich welfare, jak v individuální, tak i společenské úrovni. Může zahrnovat podporu demokratických procesů a vlády práva ve společnosti, zajištění služeb a statků ve vysoké kvalitě a za nízkou cenu, týká se ale také specifičtější role AI ztělesněné např. v podobě medicínských expertních systémů a robotů či společenských a sociálních robotů, včetně sexbotů.

Princip non-maleficence spadá do negativní oblasti morálních norem a povinností a ukládá AI, aby svým jednáním nepůsobila lidským bytostem újmu. Újmu zde musíme chápat velmi široce, od újmy na zdraví či životě až po psychickou. Vzhledem k tomu, že k lidskému welfare výraznou měrou přispívá i kvalita životního prostředí, je možné tímto principem odůvodnit morální normy zakazující AI poškozovat životní prostředí či působit újmu živočichům, zvláště těm, vůči nimž máme silné emocionální vazby.

Princip autonomie vyjadřuje vysoký respekt k lidským bytostem jako autonomním aktérům, kteří mají právo informovaně a svobodně

určovat kontury svého dobrého života a v mezích morálky tento život naplňovat. Může zakotvovat morální normy zakazující AI obelhávat lidské bytosti, nepřesně je informovat či jinak omezovat jejich možnosti informované volby, nemluvě o normách zakazujících neoprávněně je omezovat ve svobodě vyznání, názoru, projevu a pohybu.

Princip férovosti vyžaduje, aby byl vývoj, využití a regulace AI férový, nediskriminoval určité skupiny obyvatel a umožňoval jim férový přístup k benefitům – např. v oblasti vzdělání, lékařských procedur či služeb – poskytovaných AI. Princip férovosti v oblasti vzácných zdrojů – a AI a její služby zatím nejsou přístupné všem – nevyžaduje, aby všichni dostali stejně, nýbrž to, aby jejich potřeby byly férově vzaty v úvahu.

Příklad. Nemocniční tým se rozhoduje o tom, který pacient může být operovaný moderním chirurgickým robotem. Vzhledem k tomu, že mají pouze jednoho, představuje vzácný zdroj, jehož služby není možné poskytnout všem pacientům. Férovost však požaduje, aby při rozhodování o tom, který pacient bude robotem operovaný a který ne, byly v úvahu férovým způsobem vzati všichni pacienti čekající na tento typ operace.

Princip transparentnosti je specifickým principem etiky umělé inteligence a požaduje, aby systémy umělé inteligence byly kontrolovatelné a alespoň v základních obrysech srozumitelné všem lidem. Tento požadavek je klíčový vzhledem k tomu, že systémy AI jsou a stále více budou součástí naší společnosti ve všech jejích dimenzích a důvěra lidí v umělou inteligenci je nutná pro vzájemnou a dobrou koexistenci lidí a AI, bez níž se benefity s AI spojené budou jen obtížně uplatňovat. V rámci principu transparentnosti lze jako morální požadavek vyžadovat také transparentní právní úpravu regulující využití systému AI, zvláště v oblasti ochrany osobních údajů či odpovědnosti umělých systémů.

Někteří autoři formulují více základních principů (např. princip bezpečnosti, princip transparentnosti při selhání AI, princip odpovědnosti, požadavek na respekt k lidským hodnotám, princip respektu k soukromí, princip svobody, princip sdílené prosperity či princip sdílených benefitů), všechny jsou představují konkrétnější podobu výše uvedených pěti etických principů regulace AI.

CÍLE VÝZKUMU ETIKY AI

Etická regulace AI je jedním z klíčových požadavků současné doby, bez níž je koexistence systémů AI a lidských bytostí ve společnosti, včetně nezměrného množství benefitů, jež nám tato koexistence může poskytnout, nemyslitelná.

V oblasti etiky AI se nejčastěji setkáváme se dvěma etickými systémy, konsekvencialistickou etikou a deontologickou etikou. Každý ze systémů má své klady a zápory, zdá se však, že pro obecnou regulaci vývoje, využití a regulace AI je vhodnější systém deontologické etiky založený na pěti základních principech:

1. Principu beneficence.
2. Principu non-maleficence.
3. Principu autonomie.
4. Principu férovosti.
5. Principu transparentnosti.

Tyto principy umožňují formulovat celou řadu pozitivních a negativních morálních norem, jimiž by se AI měla řídit, aby se mohla stát **důvěryhodnou** (*trustworthy*) **umělou inteligencí**, která bude:

1. Respektovat lidskou důstojnost, práva a hodnoty;
2. Byla technicky natolik robustní, aby byla transparentní a spolehlivá.

Formulace těchto principů však představuje pouze první krok na cestě dlouhého hledání konkretizace těchto pravidel v praxi, zvláště v případech, kdy se AI ocitne v situaci, kdy:

1. Musí distribuovat nějakou újmu;
2. Musí distribuovat vzácné zdroje.

Příklad: Etické algoritmy autonomních vozidel se budou nevyhnutelně ocitat v situaci, kdy během kolizních situací budou muset rozhodovat o tom, jak distribuovat újmu. V takovém případě lze očekávat, že konečné řešení distribuce újmy bude zahrnovat jak deontologické požadavky určené pěti principy, tak i konsekvencialistické úvahy.

Další konkretizace etických norem, aplikovatelných na konkrétní situace každodenního života, budou zřejmě vyžadovat nejen abstraktní etické rozvahy, ale také důkladný výzkum morálních intuic lidských bytostí a kvalitní vzdělávání v oblasti etiky a informatiky.

Cíle výzkumu etické a hodnotové dimenze AI lze rozdělit do tří oblastí:

1. **Krátkodobé cíle**.
    a. Vytvoření expertní etické skupiny odborníků na normativní, aplikovanou etiku a etiku AI, která bude působit v rámci nějakého ministerstva (nejspíše Ministerstva průmyslu a obchodu). Vzhledem k jeho odbornosti by bylo vhodné, aby pověřením sestavení této skupiny bylo pověřeno Centrum Karla Čapka pro studium hodnot ve vědě a technice (www.cevast.org).

b. Vytvoření směrnice regulace výzkumu, vývoje a využití AI v České republice. Nutnost této regulace je ukotvena v samotném pojmu důvěryhodné AI a je nezbytná pro udržení kroku s vývojem ve světě a konkurenceschopnosti české vědy a techniky.

c. Vytvoření webových stránek (či podpora již existujících – viz. www.cevast.org), jejichž účelem bude edukace české populace o výhodách využívání AI a jejich bezpečnosti.

2. Střednědobé cíle.

a. Podpora vzniku doktorských (Ph.D) programů v oblasti etiky AI. Tyto doktorské programy mohou být garantovány vícero institucemi (např. Ústav státu a práva AV ČR, Ústav Informatiky AV ČR, Filosofický ústav AV ČR) a měly by sloužit k výchově odborníků, kteří, kromě výzkumu, budou moci spolupracovat s vládními orgány na implementaci etických směrnic regulujících AI a její kontrole.

b. Podpora vzniku specializovaných kurzů pro výzkumné organizace a průmyslové podniky v oblasti etiky AI, směrnic její regulace a implementace v praxi. Tyto kurzy, garantované vědeckými institucemi a Ministerstvem průmyslu a obchodu, by byly zakončeny udělením certifikátu.

3. Dlouhodobé cíle.

a. Dlouhodobé cíle jsou reprezentovány ideálem vývoje a využívání věrohodné AI a jsou podmíněny splněním cílů krátkodobých a střednědobých. Vytváření věrohodné AI je třeba zajistit na technické a netechnické úrovni.

i. Technická úroveň.

1. Etická pravidla by měla být regulačním principem již samotného vytváření návrhů systémů AI (*Ethics & Rule of Law by Design*).
2. Součástí architektury systémů AI musí být souhrn etických pravidel vymezující jejich chování.
3. Chování systémů AI musí být důkladně testováno s ohledem na jejich stabilitu, robustnost a chování v přesně definovaném rámci.

4. Systémy AI by měly ukládat jednotlivá rozhodnutí a způsoby jejich dosažení, aby bylo možné je vždy vyhledat, vyhodnotit a v případě, že rozhodnutí bylo chybné, přenastavit systém tak, aby se v budoucnosti podobných chyb vyvaroval.

ii. Netechnická úroveň.

1. Systémy AI je třeba právně regulovat, aby byla zajištěna bezpečnost jejich využití a dobře nastaveny mechanismy určování odpovědnosti v případě špatného fungování či vzniku újmy lidským bytostem (např. v případě kolize autonomních vozidel).
2. K zajištění bezpečnosti využití AI mohou sloužit také metody standardizace designu, výroby a využití v praxi.
3. Etické komise by měly na národní a podnikové úrovni sledovat využívání AI a poskytovat etické poradenství v případě vzniku konfliktních situací.
4. Česká republika by měla disponovat závaznými etickými směrnicemi regulujícími výzkum, vývoj a uplatňování AI v praxi. Tyto směrnice by se měly stát niternou součástí celého procesu vývoje AI.
5. Vzhledem k tomu, že se systémy AI stávají součástí lidské společnosti, takže lze začít hovořit o problému koexistence lidí a AI, je třeba vytvářet systémy edukace společnosti ohledně benefitů plynoucích z využívání AI a požadavků na etickou regulaci jejího chování.

ZÁVĚR

Osnova národní strategie umělé inteligence v ČR (NAIS) uvádí, že Česká republika by měla být inovačním lídrem vývoje a aplikace umělé inteligence v praxi, měla by se stát „zemí robotů". Jedním z nezbytných předpokladů splnění tohoto cíle je jasné nastavení etických pravidel regulujících chování systémů umělé inteligence takovým způsobem, aby se mohly stát součástí lidské společnosti, nepůsobily újmu, nebudily morální paniku a mohly tak plně rozvinou svůj potenciál ve všech oblastech lidského poznání a praxe, od teoretického vědeckého výzkumu přes aplikovaný až k technologickým procesům, medicínským, společenským a asistenčním robotům až po expertní systémy.

Česká republika musí při naplňování svého cíle splnit všechny předpoklady vzniku věrohodné AI, od tvorby závazných etických směrnic, přes vznik etických komisí až po nastavení podmínek výzkumu a výuky, umožňujících vznik specializovaných doktorských programů vychovávajících odborníky v oblasti etiky AI, kteří budou schopni poskytovat svou odbornou expertízu na vládní a podnikové úrovni. Jedině tak je možné zaručit naplnění cíle stát se inovačním lídrem v oblasti AI a také nastavit kontrolní mechanismy vzniku, vývoje a využívání systémů AI v naší společnosti.