Etické principy regulace chování umělé inteligence

David Černý

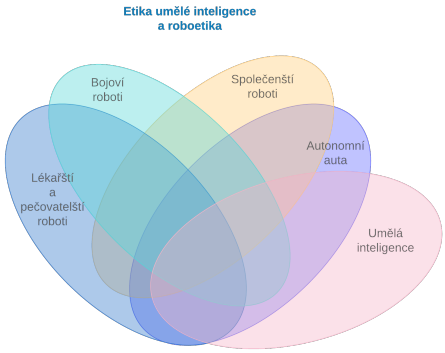
Ústav státu a práva AV ČR, v. v. i.
Národní 18
11600 Praha 1
Centrum Karla Čapka pro studium hodnot ve vědě a technice
www.cevast.org
Email: david.cerny@ilaw.cas.cz

Abstrakt

V tomto příspěvku se pokusím načrtnout kontury dvou důležitých etických teorií, jež nacházejí významné uplatnění v současné diskusi etiky umělé inteligence a roboetiky: utilitarismu a deontologie. Předložím stručnou charakteristiku obou etických teorií a poukáži na jejich silné a slabé stránky při řešení praktických problémů.

1 Etika a umělá inteligence

Vzhledem k mohutným pokrokům na poli výzkumu a aplikace umělé inteligence a robotiky se do popředí zájmu dostávají také hodnotové a etické otázky, které se dotýkají výzkumu, návrhu designu, pravidel jednání a nasazení systémů AI. Podívejme se na následující obrázek.



Obr. 1: Etika AI a roboetika

Obrázek ukazuje, že oblasti etického zkoumání AI a jejích různých forem „ztělesnění" se překrývají, vyžadují však také řešení vlastních specifických problémů. Např. v případě etiky společenských robotů se klade velký důraz na sociální dimenzi těchto strojů, která se promítá nejen do hledání etických pravidel regulace chování, ale výraznou měrou zasahuje do samotného procesu vytváření jejich designu, který musí splňovat charakteristiky nutné pro to, aby tito roboti mohli být skutečnými lidskými společníky. U bojových robotů se střetáváme s klasickými problémy *ius in bello* (etická pravidla vedení válečných konfliktů), u autonomních vozidel nás zase zajímají etická pravidla, jimiž by se tyto automobily měly řídit v případě neodvratné kolize, v níž je možné rozhodnout o distribuci újmy.

V tomto krátkém příspěvku se nemohu věnovat obecným otázkám etiky umělé inteligence či specifickým etickým problémům v rámci jejích různých sfér aplikace. Pokusím se proto charakterizovat dva důležité etické systémy, s nimiž se v dnešní diskusi setkáváme nejčastěji (utilitarismus a deontologickou etiku), a ukázat jejich silné i slabé stránky.

Oba dva etické systémy mají celou řadu různých variant, kterým se zde nemohu věnovat. Pokusím se je proto popsat způsobem, jenž se mi zdá nejvhodnější pro jejich aplikaci v kontextu AI a robotiky. Je však docela dobře možné, že různá ztělesnění AI budou využívat různé etické přístupy, například kombinaci utilitarismu a základních deontologických omezení, pravděpodobně se také setkáme s tím, že v rámci odlišných oblastí využívání AI bude třeba klást důraz na jiné etické systémy a jejich kombinace. Autonomní vozidla se například budou řídit sdílenou etikou AI, ale v případě kolizních situací se do popředí dostane utilitaristická etika minimalizace újmy. Podobně v případě lékařských expertních systémů bude důraz kladený na etiku deontologickou, v některých případech doplněnou utilitarismem, zatímco společenští roboti se budou podřizovat některým principům etiky péče a etiky ctností.

2 Utilitarismus

2.1 Charakterizace utilitarismu

Utilitarismus představuje jednu z nejvlivnějších současných etických teorií s širokou škálou aplikací ve všech oblastech lidské praxe, od vědeckého výzkumu, medicíny, vojenské etiky až po populační etiku či etiku umělé inteligence a roboetiku. Jedním z důvodů široké rozšířenosti utilitarismu je zdánlivě snadná aplikovatelnost v praxi, včetně řešení složitých etických dilemat, a poměrně minimální metaetické předpoklady.

Utilitarismus vychází z několika jednoduchých předpokladů (Mulgan (2007).

1. **Konsekvencialismus**. Je-li X nějaký aktér, ϕ_i možné jednání v situaci S a $\delta_1, \ldots, \delta_n$ důsledky ϕ_i v S, potom o morálním hodnocení (správné – nesprávné) ϕ_i rozhodují pouze důsledky $\delta_1, \ldots, \delta_n$(Driver (2012)).

2. **Nestrannost.** Je-li X nějaký aktér a $\phi_1, \ldots, \phi_n$ jsou možné alternativy jednání v situaci S, potom X musí důsledky $\phi_1, \ldots, \phi_n$ posuzovat nestranně.

Stručný komentář:

1. X se v S rozhoduje, zda porušit (ϕ) či neporušit ($\neg\phi$) slib. Ve své praktické rozvaze bere v úvahu pouze jediný normativní faktor udělující ϕ a $\neg\phi$ jejich etické hodnocení: důsledky. Do své úvahy tedy nezačleňuje jiné možné normativní faktory, např. to, že porušení slibu může samo o sobě představovat typ jednání, které je morálně nesprávné. Jsou-li důsledky ϕ „lepší“ než důsledky $\neg\phi$, potom X slib poruší.

2. X se v situaci S rozhoduje mezi jednáním ϕ_1 a ϕ_2, jejichž důsledky se dotýkají dvou lidských bytostí: ϕ_1 se dotýká L_1, ϕ_2 se dotýká L_2. X musí důsledky obou variant jednání hodnotit bez ohledu na to, zda a případně jak je vztažený k L_1 a L_2 (např. příbuzensky), dokonce i bez ohledu na to, zda platí $X = L_i$.

Kromě dvou předpokladů vymezujících třídu morálně relevantních faktorů (konsekvencialismus) a způsobu jejich hodnocení (nestrannost) ještě potřebujeme procesní pravidlo, které pro každého aktéra X, situaci S a možné varianty jednání $\phi_1, \ldots, \phi_n$ jednoznačně vybere to jednání ϕ_i, které je morálně správné.

Předpokládejme, že se určitý systém umělé inteligence (SUI) nachází v situaci S, v níž se před ním otevírá několik možností jednání $\phi_1, \ldots, \phi_n$. SUI přiřadí každé variantě jednání ϕ_i určitý užitek I_i. Předpokládejme dále, že třída hodnot $\{I_1, \ldots, I_n\}$ má maximum, jímž je řekněme hodnota I_3. Podle utilitarismu je pro SUI v situaci S, zahrnující varianty jednání $\phi_1, \ldots, \phi_n$, morálně správné jednání, které maximalizuje užitek, tj. v našem případě ϕ_3. Namísto užitku budu také hovořit o hodnotě jednání ϕ_i a označovat ji $|\phi_i|$. Je zřejmé, že $|\phi_i| = I_i$.

Utilitarismus tedy požaduje, abychom **maximalizovali užitek**; maximalizace užitku je jediným procesním pravidlem tohoto etického systému.

Hodnotu $|\phi_i|$ jednání ϕ_i jsem označil za užitek. Utilitarismus je nekompletní, dokud přesně nespecifikujeme povahu tohoto užitku. Jak však užitek definovat a kvantifikovat? Utilitarismus patří mezi tzv. welfaristické teorie, považuje welfare (jak dobrý život určitého subjektu S je) za klíčový normativní faktor, neboť důsledky $\delta_1, \ldots, \delta_n$ variant jednání $\phi_1, \ldots, \phi_n$ vyhodnocujeme na základě jejich pozitivního či negativního (případně i neutrálního, nastane-li taková situace) příspěvku k welfare S.

Užitek je možné definovat různými způsoby, které odpovídají různým koncepcím welfare. Nejčastěji se v kontextu utilitarismu setkáváme s dvěma teoriemi welfare (dobrého života):

1. **Hédonismus** (Carson (2011)).

   - Věcný stav „S zažívá v čase t_i potěšení v míře D“ je jedinou kladnou hodnotou (pozitivním příspěvkem k dobrému životu).
   - Věcný stav „S zažívá v čase t_i utrpení v míře D“ je jedinou negativní hodnotou (negativním příspěvkem k dobrému životu.

2. **Preferencialismus** (Olsaretti (2006)).

   - Věcný stav „S v čase t_i touží po tom (přeje si, aby), aby propozice P byla pravdivá v míře D a P je pravdivá“ je jedinou kladnou hodnotou (pozitivním příspěvkem k dobrému životu).
   - Věcný stav „S v čase t_i touží po tom (přeje si, aby), aby propozice P byla pravdivá v míře D a P je nepravdivá“ je jedinou negativní hodnotou (negativním příspěvkem k dobrému životu.

Velmi důležitým prvkem utilitaristického kalkulu je předpoklad, že důsledky jednání lze vyjádřit v pojmech užitku, tj. lze je kvantifikovat a srovnávat (maximalizace totiž vyjadřuje ideu, že morálně správné jednání je takové, které má nejvyšší hodnotu). Jak by taková kvantifikace mohla vypadat? Ukážeme si to na příkladu hédonismu (Feldman (2004)). Nejdříve musíme stanovit nějakou základní jednotku potěšení – nazveme si ji hédonem a označíme H –, dále musíme určit základní jednotku utrpení – nazveme si ji dolorem a označíme D –, a nakonec určíme, že platí:

$$|H| = |D|. \tag{1}$$

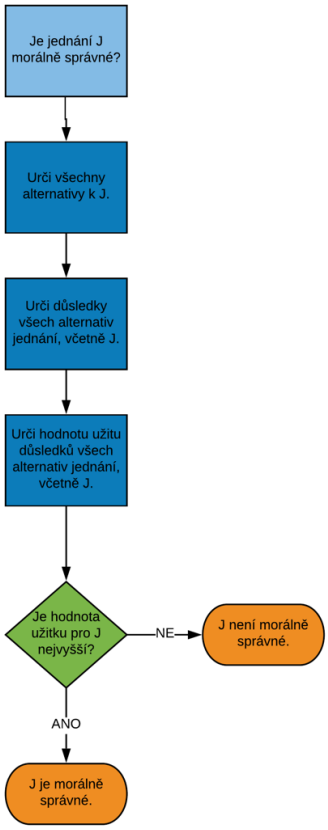
Absolutní hodnota hédonu je rovna absolutní hodnotě doloru, dolory mají jako negativní hodnoty zápornou hodnotu. Dále určíme, že množství hédonů nějaké pozitivního věcného stavu v t_i je rovno D, tj. míře potěšení v t_i, případně že množství hédonů $Sum(D)_{<t_i,t_j>}$ věcného stavu v intervalu $< t_i, t_j >$ je rovno celkovému množství hédonů mezi t_i a t_j.

Analogicky budeme postupovat v případě dolorů. Konečným krokem je určení balance hédonů a dolorů, vyjádřené pojmem hédono-dolorické balance (HDB), pro niž platí:

$$HDB = H + D. \quad (2)$$

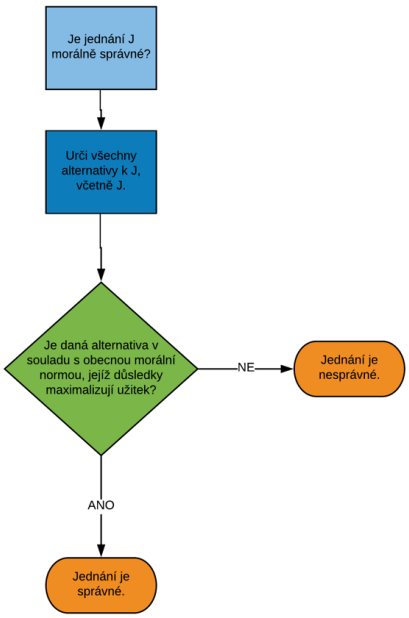
Řekněme, že HDB osoby S je rovna a a aktér X zvažuje varianty jednání ϕ_1, ϕ_2, ϕ_3, pro které (vzhledem k S) platí: $|\phi_1| = 5$, $|\phi_2| = -7$, $|\phi_3| = 12$. Tato tři čísla určují, jak (kvalita – pozitivně – negativně) a v jaké míře (kvantita) jednání ϕ_1, ϕ_2, ϕ_3 ovlivní HDB S. Je zřejmé, že dvě varianty – ϕ_1, ϕ_3 – zasáhnou do welfare S pozitivně (zlepší život S), zatímco jedna varianta – ϕ_2 – ovlivní welfare S negativně. Vzhledem k tomu, že hodnota jednání ϕ_3 je nejvyšší, jedná se o morálně správné jednání.[1]

Celou rozhodovací proceduru názorně shrnuje následující obrázek:



Obr. 2: Utilitarismus činů: rozhodovací procedura

Utilitaristická praktická rozvaha může mít ještě jednu podobu. Namísto hodnocení důsledků každého konkrétního činu (proto o výše popsané verzi utilitarismu hovoříme jako o **utilitarismu činů**) hodnotíme obecné morální normy a volíme takové, jejichž internalizace ve společnosti má nejlepší důsledky (Hooker (2000)). Je-li tedy Δ obecná morální norma, pod níž spadající jednání má nejlepší možné důsledky, potom pro aktéra X v situaci S platí, že spadá-li z existující množiny jednání $\{\phi_1, \ldots, \phi_n\}$ nějaká varianta ϕ_i pod normu Δ, potom je ϕ_i morálně správné jednání. Této verzi utilitarismu říkáme **utilitarismus pravidel** a jeho rozhodovací proceduru shrnuje následující obrázek:



Obr. 3: Utilitarismus pravidel: rozhodovací procedura

2.2 Aplikace utilitarismu v etice AI a roboetice

Utilitarismus představuje velmi přímočarou etickou teorii, která může být poměrně snadno aplikovatelná v praxi. Přesto jsou s ní spojeny určité potíže, které se zde pokusím velmi stručně popsat.

Jedním z důležitých předpokladů vtělených do systému utilitaristickém etiky je nestrannost: jestliže aktér X v dané situaci S hodnotí možné důsledky svého jednání (variant jednání, přičemž i nejednání je variantou jednání), musí je zvažovat bez ohledu na to, zda se tyto důsledky dotýkají jeho samého, lidí, s nimiž je spřízněný či spojený různými společenskými vazbami, jako je přátelství, sousedství apod. Mohl by se ale takovou etikou řídit např. pečovatelský robot, který je naším majetkem a byl zakoupený pro to, aby se staral o naši babičku v domově důchodců? Chápali bychom jako eticky přijatelné, aby tento robot nepřetržitě monitoroval své okolí, další obyvatele domova důchodců a neustále na základě maximalizace užitku hodnotil, komu má ve skutečnosti v dané situaci věnovat svou pozornost?

Tento problém by zřejmě šlo vyřešit tak, že by se tito roboti řídili pravidlem, podle kterého mají vždy a za všech okolností dávat přednost svému majiteli či tomu, do jehož péče jsou svěřeni. V takovém případě by se však nejednalo o utilitaristickou etiku, přinejmenším

[1] Užitek $|\phi_i|$ jednání ϕ_i je odvozenou veličinou, kterou získáme tak, že určíme množinu $\{\delta_1, \ldots, \delta_n\}$ důsledků ϕ_i, každému důsledku přiřadíme jeho užitek a posléze je sečteme.

ne o její čistou verzi, protože by byla doplněna tímto deontologickým pravidlem, v jehož rámci (vždy dávat přednost majiteli) by robot dále mohl jednat ve shodě s utilitarismem.

Uvažme jiný příklad, etiku kolizních situací autonomních vozidel (AV). Tato etika může mít poměrně jednoduchou strukturu: za všech okolností minimalizuj újmu. Znamená to tedy, že se AV mají touto etikou řídit vždy, i za tu cenu, že obětují posádku v případě, že užitek jejího obětování je byť i jen o málo větší, než užitek neobětování? Budou si však lidé ochotni si taková AV kupovat, budou ochotní jim svěřit své blízké a milované? Pokud ne, bude to mít důležité etické důsledky. Autonomní doprava slibuje snížit počet obětí dopravních nehod až o 93 %, nicméně nebudou-li lidé ochotní si tato utilitaristická vozidla kupovat, výraznou měrou to zbrzdí nástup autonomní dopravy se všemi benefity, které slibuje. Podobně si ale lidé nebudou přát vozidla, které dají přednost posádce za všech okolností. Například v situaci jeden řidič – pět dětí na přechodu si většina lidí bude myslet, že by AV mělo obětovat řidiče, tedy posádku.

Opět by bylo možné tuto situaci vyřešit tak, že by AV kombinovala dva typy etických úvah, utilitaristické a deontologické. Prostřednictvím experimentálního výzkumu morálních intuic potenciálních uživatelů AV by bylo možné zjistit, kde nastavit prahy, oddělující oba dva etické přístupy. Výsledkem by byl hybridní etický systém, který by AV přikazoval dát přednost posádce bez ohledu na důsledky (deontologie), ale pouze do určité poměru způsobené újmy. Za tímto empiricky zjištěným prahem (např. poměr 1 život ku 3) by se AV již chovala utilitaristicky. Tuto hybridní povahu etiky AV lze ospravedlnit vnějších kritériem utilitaristické povahy, totiž tím, že zavedení AV do provozu ušetří ročně desítky tisíc životů, obrovské škody na zdraví a na majetku.

I kdybychom však došli k nějakému hybridnímu etickému systému AV, problémy by nastaly v okamžiku překročení prahu směrem k utilitaristické komponentě. Jak budeme určovat důsledky jednání? Nabízí se jednoduché řešení: určíme tři kategorie újmy – újmu na životech, újmu na zdraví a újmu na majetku. Provádět kvantifikaci a srovnávání v rámci jednotlivých kategorií by nemuselo být tak obtížné (1 život vůči pěti, jedna zlomená ruka oproti bouli na hlavě, škoda 100 000 korun vůči miliónové škodě apod.), jak však budeme provádět kvantifikaci a srovnávání mezi různými kategoriemi újmy, např. mezi újmou na životech a újmou na zdraví? Je možné říci, že jeden lidský život je převážený pěti zlomenými rukami a jednou zlomeninou páteře? A skutečně je eticky korektní taková srovnávání provádět a dovolit AV, aby dalo přednost minimalizaci obrovské majetkové újmy a obětovalo jeden lidský život?

Nebo se zamysleme nad lékařskými expertními systémy, které by mohly lékařům pomáhat se svízelnými etickými problémy, např. s alokací vzácných zdrojů v kontextu transplantační medicíny. Představme si, že lékaři mají pět pacientů, kteří nutně potřebují transplantaci vitálního orgánu, nemají však žádného vhodného dárce. Může expertní systém doporučit zabití zdravého pacienta bez příbuzných a použití jeho orgánů? Z hlediska utilitarismu není rozdíl mezi zabitím (usmrcením pacienta) a ponecháním zemřít (neposkytnutím orgánů pěti pacientům) – důsledek je pokaždé stejný. Chtěli bychom však, aby se lékaři řídili radami takto uvažujícího systému umělé inteligence? Neměl by obsahovat nějaká zásadní omezení (spadající opět do deontologické etiky) typu „usmrcení nevinného člověka je za všech okolností nepřípustné, bez ohledu na důsledky“? Výsledkem by opět byl systém, který je z etického hlediska hybridní, neboť kombinuje dva typy úvah - utilitaristickou a deontologickou.

Na základě těchto úvah, které zde nemohu dále rozvádět, se domnívám, že se s utilitaristickou etikou nebudeme v čisté podobě v oblasti AI a robotiky příliš setkávat, třebaže má celou řadu velmi vlivných zastánců a po překonání určitých problémů (např. určení hodnot jednání) by byla velmi snadno implementovatelná do řídících algoritmů systémů umělé inteligence.

3 Deontologická etika

Pod hlavičkou „deontologické systémy“ se skrývá celá řada etických přístupů, které je obtížné jednoduše a jednoznačně charakterizovat (Kagan (1998)). 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.

Deontologické etiky tedy mohou postupovat tak, že nejdříve určí obecné morální principy, v nichž jsou zakotveny obecné morální normy. Nechť Υ představuje třídu morálních norem. Třída $\Phi, \Phi \subset \Upsilon$ je třídou pozitivních morálních norem vyjádřených formou příkazů či doporučení („Musíš pomáhat rodině“, „Měl bys pomáhat hladovějícím dětem v Africe“ apod.), zatímco třída $\Psi, \Psi \subset \Upsilon$ je třídou negativních morálních norem („Nikdy nesmíš usmrtit nevinnou lidskou bytost“, „Mučení je za všech okolností nepřípustné“ apod.). Pro tyto tři třídy platí:

$$\Phi \cup \Psi = \Upsilon \tag{3}$$

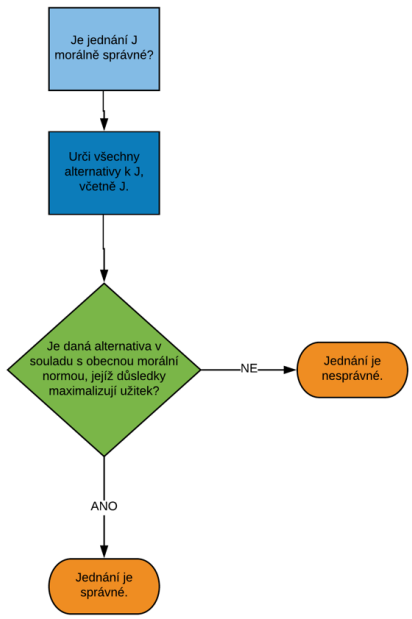
$$\Phi \cap \Psi = \emptyset \tag{4}$$

Třídy Φ a Ψ se neliší pouze způsobem ukládání závazku, ale také tím, že o třídě Φ často říkáme, že její prvky jsou platné *semper sed non ad semper* (vždy, ne za všech okolností), zatímco prvky třídy Ψ jsou platné *semper et ad semper* (vždy a za všech okolností) (Finnis (1991)). Je to poměrně jednoduché: negativní normy jasně vymezují určité formy jednání a specifikují je jako morálně nepřípustné, pozitivní normy sice platí vždy, je však možné je naplňovat v různých situacích různým způsobem. Zákaz vraždy jasně definuje, co nesmíme, a platí vždy a za všech okolností, příkaz pomáhat bližnímu však nespecifikuje možné varianty jednání, které lze chápat jako pomoc bližnímu, ani neurčuje, koho bližním máme vlastně myslet (pomoc bližnímu se tak může za určitých okolností prostě redukovat na pomoc vlastní rodině).

Deontologická etika tedy požaduje, abychom nejdříve určili povahu jednání ϕ_i (vražda, zabití v nedbalosti, krádež, pomoc bližnímu, nevěra, sebevražda…) a poté povahu ϕ_i srovnali s Υ:

- Spadá-li ϕ_i pod Φ, potom je ϕ_i morálně správné a *ipso facto* přípustné.
- Spadá-li ϕ_i pod Ψ, potom je ϕ_i morálně nepřípustné.
- Nespadá-li ϕ_i pod Υ, potom je morálně indiferentní.

Rozhodovací proceduru v rámci deontologické etiky shrnuje následující obrázek:



Obr. 4: Deontologická etika: rozhodovací procedura

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“.

Deontologická etika se často interpretuje tak, že ukládá negativní a pozitivní povinnosti:

1. **Negativní morální povinnost**. Je-li X nějaký aktér, potom X má negativní povinnost $P^n_{\phi_i}$, existuje-li jednání ϕ_i takové, že $P_{\phi_i} \in \Psi$.
2. **Pozitivní morální povinnost**. Je-li X nějaký aktér, potom X má pozitivní povinnost $P^p_{\phi_i}$, existuje-li jednání ϕ_i takové, že $P^p_{\phi_i} \in \Phi$.

Pozitivní a negativní morální povinnosti nelze mezi sebou nějakým způsobem poměřovat, tj. jakási morální aritmetika, s níž jsme se setkali např. v případě etiky minimalizace újmy (3 životy vůči jednomu apod.), je možná pouze v rámci jednotlivé kategorie morálních povinností. Není tedy např. možné naplňovat pozitivní povinnosti (např. zachránit život pěti pacientům), pokud by byla porušena negativní povinnost (neusmrtit nevinnou lidskou bytost).

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 utilitaristické 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:

Obr. 5: Principy etiky AI

Podívejme se na jednotlivé principy poněkud podrobněji:

- **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 spravedlnosti** 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.
- **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.

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:

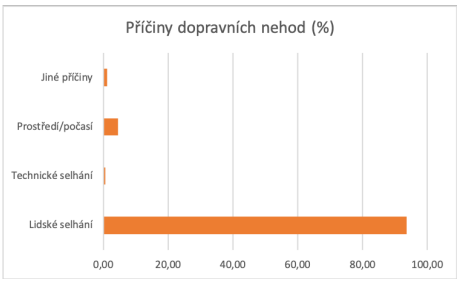
1. Respektovat lidskou důstojnost, práva a hodnoty.
2. Byla technicky natolik robustní, aby byla transparentní a spolehlivá.

I tento přístup má však celou řadu problémů, jež nastávají v okamžiku, kdy se jednotlivé principy dostávají do konfliktu. Chápeme-li je jako absolutní, neexistuje žádné řešení tohoto konfliktu a umělá inteligence řídící se tímto přístupem by se v praxi velmi často dostávala do situace, kdy by nevěděla, jak postupovat. Řešení by mohlo spočívat v tom, že bychom principy a z nich odvozené normy a povinnosti nechápali absolutně, jako platné vždy a všude, ale jako tzv. *prima facie* principy (normy a povinnosti). O nějakém principu (morální normě) říkáme, že je *prima facie*, pokud je platný, mohou ale nastat situace poskytující důvody převažující platnost tohoto principu. Otázkou potom samozřejmě je, jak přesně určit, kdy a které důvody daný princip převažují.

4 Case study: autonomní vozidla

Domnívám se, že popsané problémy poměrně přesvědčivě ukazují, že se v případě etiky umělé inteligence budeme muset vzdát představy, že k její regulaci bude možné využít pouze jeden „čistý“ etický systém (jako je utilitarismus či nějaká verze deontologické etiky). Spíše budeme muset využívat více etických systémů, každý zaměřený na specifické cíle daného systému umělé inteligence.

Zaměřme se např. na etiku regulující provoz autonomních vozidel (AV). V případě AV musíme rozlišit několik etických rovin; první z nich se týká samotné existence těchto sofistikovaných robotů a jejich etických a společenských důsledků. Podívejme se na následující graf (Mauer M. (2015)):



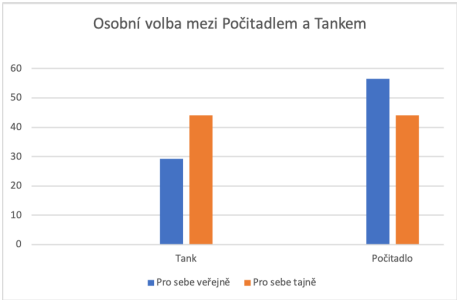
Obr. 6: Příčiny dopravních nehod

Drtivá většina dopravních nehod je způsobena lidským faktorem. Na celém světě dochází ročně zhruba k 1.240.000 smrtelných dopravních nehod, v ČR bylo v roce 2017 při nehodách usmrceno 502 lidí. Všechna tato čísla ukazují, že zavedení AV do provozu je skutečným morálním imperativem, protože bude mít blahodárné účinky, výraznou měrou sníží počet dopravních nehod (v ideálním případě až o 93 %). Máme zde tedy první morálně relevantní faktor: důsledky zavedení určité technologie, v tomto případě AV, do praxe. Když

zvažujeme etická pravidla regulace chování AV, musíme mít tento faktor neustále na paměti. Je nepochybné, že se AV budou dostávat do eticky problematických situací a někdy v nich budou jednat v rozporu s morálními intuicemi určité části populacem, tento fakt je však vyvážen benefity, plynoucími z existence AV.

Aby však celá společnost mohla mít prospěch z AV, je nezbytné, aby se stala skutečně rozšířenou technologií, což samozřejmě také předpokládá, že si je lidé budou chtít kupovat, budou ochotní jim svěřovat své životy a životy a životy svých blízkých. První normativní faktor (prospěch z existence a využívání AV) představuje silné – zřejmě dokonce přesvědčivé – důvody k zavedení AV do provozu, to ale předpokládá ochotu potenciálních kupců si tato vozidla skutečně kupovat. Je proto třeba vědět, jaké jsou morální intuice těchto potenciálních uživatelů AV, zvláště v případě možných kolizních situací.

Výzkumy ve světě (Bonnefon (2016)) i náš vlastní, zatím nepublikovaný výzkum ukazují, že lidé nemají konzistentní postoje. Třebaže si většina z nich myslí, že správný morální algoritmus řídící AV by měl respektovat principy etiky minimalizace újmy (utilitarismus) – takovému autu jsme dali název Počtář –, pokud by si mohli volit z více variant, dali by pro sebe a svou rodinu přednost typu Tank, tj. algoritmu, který by v kolizních situacích chránil posádku. Např. v případě, že by volili pro sebe, rozhodovali by se v tajné volbě (nebylo by zřejmé, zda vlastní Tank, nebo Počtáře) výrazně jinak, než ve volbě veřejné či v obecném hodnocení, jaký typ morálního algoritmu je správný:



Obr. 7: Tajná a veřejná volba řídícího algoritmu

Pokud tedy chceme, aby AV naplnila svůj morální potenciál, musíme najít rovnováhu mezi etickými intuicemi jejich potenciálních uživatelů a regulací chování AV v kolizních situacích. Software AV zřejmě bude využívat dva etické systémy – utilitarismus, ukládající minimalizaci újmy –, a nějakou formu deontologické etiky, ukládající ochránit posádku. O aplikaci prvního či druhého systému bude rozhodovat metaprincip („nyní minimalizuj újmu“, „nyní chraň posádku“), který v případě překročení empiricky (na základě výzkumu morálních intuic) určeného prahu založeného na srovnávání celkové újmy rozhodne, zda v dané konkrétní situaci použít první či druhý etický systém.

V praxi by to mohlo vypadat např. takto: práh (pro životy) 1 : 2. Je-li tedy v AV jeden řidič a na silnici jeden či dva chodci, metapravidlo určí, že v případě hrozící smrtelné kolize se AV má řídit deontologickou etikou. Je-li práh překročený (jeden řidič – tři chodci, dva spolujezdci – šest chodců . . .), metapravidlo rozhodne, že se AV musí řídit etikou minimalizace újmy.

Poděkování

Tento příspěvek vznikl s podporou projektu TAČR TL01000467 s názvem *Etika provozu autonomních vozidel*. Děkuji Jiřímu Wiedermannovi za komentáře k první verzi tohoto textu.

Reference

Anderson, M. a Anderson, S. L. (zost.) (2011). *Machine Ethics*. Cambridge University Press, 1. vyd.

Bonnefon, J-F., S. A. R. I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–76.

Carson, T. L. (2011). *Value and the Good Life*. University of Notre Dame.

Driver, J. (2012). *Consequentialism*. Routledge.

Feldman, F. (2004). *Pleasure and the Good Life. Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford University Press.

Finnis, J. M. (1991). *Moral Absolutes. Tradition, Revision, and Truth*. The Catholic University of America Press.

Hooker, B. (2000). *Ideal Code, Real World*. Oxford University Press.

Kagan, S. (1998). *Normative Ethics*. Routledge.

Lin, P., Abney, K. a Bekey, G. A. (zost.) (2014). *Robot Ethics. The Ethical and Social Implications of Robotics*. MIT Press.

Lin, P., Jenkins, R. a Abney, K. (zost.) (2017). *Robot Ethics 2.0 From Autonomous Cars to Artificial Intelligence*. Oxford University Press.

Mauer M., Gerdes, C. L. B. W. H. (zost.) (2015). *Autonomous Driving. Technical, Legal and Social Aspects*. Springer.

Mulgan, T. (2007). *Understanding Utilitarianism*. Acumen.

Olsaretti, S. (zost.) (2006). *Preferences and Well-Being*. Cambridge University Press.