Esej

Morální problémy autonomních vozidel[1]

Robin Kopecký
Centrum Karla Čapka pro výzkum hodnot ve vědě a technice, Praha
Přírodovědecká fakulta Univerzity Karlovy, Praha
robin.kopecky@natur.cuni.cz

1. Úvod

Co jsou autonomní vozidla (AV)? Technická terminologie rozlišuje celou řadu kategorií,[2] pro potřeby filosofické eseje budou stačit následující: Asistent řidiči pomůže například s parkováním nebo udržováním rychlosti. Automat je již schopen sám dojet z místa A na místo B. A nakonec, autonomní vůz umí nejen vykonat vše předcházející, ale je navíc vybaven aktivními prvky: kamerami, radarem a jinými senzory, a hlavně komunikuje – vyměňuje si informace s dalšími účastníky provozu. Budeme dále jako AV souhrnně (zjednodušeně) označovat takový silniční dopravní prostředek, který je vybaven samo-řídicím softwarem, jenž je nezávislý na posádce.

Jaký je účel AV? Slouží k dopravě lidí a věcí. Pointa je v tom, že AV to po všech stranách zvládnou lépe než stávající kombinace auta a lidského řidiče.

1 Tento příspěvek vznikl s podporou Grantové agentury Univerzity Karlovy v rámci projektu GA UK č. 929216: „Faktory ovlivňující altruismus".

2 Dle mezinárodní standardizace: SAE On-Road Automated Vehicle Standards Committee et al., Taxonomy and Definitions for Terms Related to On-road Motor Vehicle Automated Driving Systems. *SAE Standard J3016*, 01–16, 2014.

Ukažme si to na příkladu dopravních nehod v EU.[3] I když se investují značné prostředky do zvýšení bezpečnosti na evropských silnicích, ročně na nich zemře 25–30 tisíc lidí a okolo 130 tisíc se zraní. Vidíme, že problém leží přesně mezi volantem a sedadlem: je jím řidič. V USA[4] je 94 % nehod zapříčiněno chybou řidiče a k polovině z nich dojde na dálnici, tedy ve snadné dopravní situaci. Nahradíme-li řidiče něčím (nebo někým?), co/kdo bude dělat méně chyb, jen v EU tím každoročně zachráníme kolem 30 tisíc lidí a předejdeme tisícům zraněných. Řídicí software – na rozdíl od řidiče – nejezdí pod vlivem drog, nepíše textové zprávy a neusíná. Zavedení AV by kromě toho, že ubyde mrtvých, přineslo další nesporná pozitiva: synchronizovaný provoz AV sníží emise a zrychlí dopravu. Lidé příliš mladí na to, aby sami řídili, nebo lidé se zdravotním hendikepem získají možnost větší mobility. Řidiči nebudou ztrácet čas a nervy řízením, namísto toho mohou během cesty spát nebo užívat AV jako svou mobilní pracovní kancelář. AV nemusí nezbytně vést ani ke zvýšení počtu automobilů, ani k poklesu využívání hromadné dopravy. Lidé nebudou muset vlastnit auto, zvláště když jej využívají jen několik procent dne, většinou na dojíždění do práce. Sdílení aut bude pohodlné – auta se budou sama půjčovat i vracet. Městská hromadná doprava nebude limitována počtem zkušených řidičů. Vidíme jen samé výhody. Měli bychom proto AV jako racionální volbu zavést všude a co nejdříve? Technici jsou skvělí v řešení problémů našeho světa, činí život kvalitnějším – v této eseji se budu ale snažit dostát roli filosofa, totiž ke každému (technickému) řešení, se kterým přijde inženýr, dosadit další (filosofický) problém.

2. Filosofické problémy techniky minulé, brzy současné a budoucí

Dříve než obrátíme pozornost k hlavnímu problému – autonomním vozidlům, zasadíme tento problém do historického kontextu. V průběhu minulých dvou staletí docházelo díky technologickému pokroku v každodenním životě nejen ke kvantitativním změnám, kdy byly například olejové lampy hromadně nahrazeny kvalitnějším elektrickým osvětlením, ale také ke změnám kvalitativním, nejčastěji spočívajícím v náhradě živého aktéra strojem. Například namísto aktéra–koně se stal původcem pohybu parní stroj nebo spalovací motor. Živý aktér v určité podobě – strojvůdce ovšem nadále zů-

3 Údaj ze zprávy *Evropské komise* [online]. Dostupné na: https://ec.europa.eu/transport/road_safety/specialist/statistics_en; [cit. 22. 9. 2018]. Souborná data Eurostatu [online]. Dostupné na: http://ec.europa.eu/eurostat/statistics-explained/index.php/Road_safety_statistics_at_regional_level; [cit. 22. 9. 2018].

4 Česká data o nehodovosti zpracovává Centrum dopravního výzkumu [online], přehledné shrnutí příčin je dostupné na: http://www.czrso.cz/clanky/hloubkova-analyza-silnicnich-dopravnich-nehod-hlavni-priciny-vzniku-nehod/; [cit. 22. 9. 2018].

stával na dohled pasažérům, sice už nebyl původcem pohybu, ale měl nad dějem alespoň kontrolu. Jakmile došlo k eliminaci i tohoto aktéra, například ve výtazích, vzbudila jeho absence mezi veřejností nejistotu. Strach z mechanismu fungujícího bez živého aktéra se zde projevil ve své nejprostší podobě.

Dalším příkladem revoluční kvantitativní změny byla automatizace výroby, která způsobila nejen redukci, ale i úplné zrušení některých pracovních pozic. Jako příklad uveďme pozici „počítače", která spočívala ve vykonávání pomocných numerických výpočtů. „Počítačem" byl člověk jen se svou (rozšířenou) myslí, tužkou, papírem, tabulkami a pravítkem. Během první světové války počítali lidské počítače trajektorie pro artilerii, což je úloha, kterou bylo možné koncipovat lidskou myslí. Počítající si mohli představit letící projektil a hlavně věděli, co a proč počítají. Ve velkolepém projektu „Manhattan" byly během druhé světové války zaměstnány stovky takových lidských počítačů – jejichž pozice byla ovšem diametrálně odlišná. Nejenže si nijak nemohli představit formální matematické postupy, ale dokonce ani nesměli vědět, proč takové výpočty dělají a k čemu budou následně sloužit. Dnes pojmenování počítač plně náleží stroji složenému z křemíkového *hardwaru*, tedy už nikoliv člověku z *wetwaru*, z masa a kostí. Je tomu tak kvůli rychlosti a přesnosti a nemá ani smysl tyto dva typy „počítačů" srovnávat. Na těchto dvou jednoduchých příkladech je patrné, že došlo k náhradě prosté mechanické – fyzické na úrovni pohybu, ale i intelektuální a kreativní na kognitivní úrovni, přičemž obě změny byly kvalitativní.

Mezi prosté problémy současnosti a blízké budoucnosti můžeme zařadit skutečnost, že jak v činnostech, které vyžadují „strojovou" přesnost, rychlost apod., tak i v činnostech kreativních člověk prohrává v přímém souboji se stroji. Již v roce 2006 porazil počítač světového velmistra v šachu Garriho Kasparova. Za předpokladu, že se náhrada stroji stane levnější a produktivnější alternativou využití člověka, je na místě se v západním světě obávat o pracovní uplatnění lidí s nízkou odborností. Část lidstva by se mohla snadno stát na trhu práce nezaměstnatelnou. Jistou analogii vidím v historii armád. Dnes již není třeba *„kanónenfutru"* z řad poddaných, protože současní „pěšáci" v poli obsluhují specializovanou techniku – pro nespecializované osoby tedy není využití a nemá smysl je masově nabírat do armády. Zmíněné problémy zastarávání a neefektivity člověka se už staly jakýmsi klišé debat o automatizaci. Je možné, že daný problém zmírní nebo vyřeší neviditelná ruka trhu. Mezi předpokládané scénáře patří přesunutí volných pracovníků z výroby do služeb, zkrácení pracovní doby a možná i zavedení nepodmíněného příjmu zajišťujícího základní potřeby těch nejméně uplatnitelných.

To, co přináší zásadní kvalitativní změny, a to, co je revoluční a prakticky nevratné, souborně nazývám *„superproblémy budoucnosti"*. Prvním z nich je *„superdostupnost"* technologií. Důsledkem snižování nákladů na výro-

bu efektivních strojů a zdvojnásobení výkonu výpočetní kapacity bude dle Mooreova zákona[5] již za 18 měsíců větší dostupnost výkonných technologií. Poté bude teoreticky možné, aby si každý doma v garáži vyvinul nový vzduchem šiřitelný smrtelný virus nebo si sestrojil kufříkovou atomovou bombu. Dokud byly prostředky s potenciálně apokalyptickými účinky v rukou jen malého počtu mocenských skupin – států, bylo možné předpokládat, že zastanou role racionálních hráčů. V případě, že se taková moc dostane do rukou téměř kohokoliv, bude na základě zákona velkých čísel téměř jisté, že bude *„superdostupná"* technologie zneužita šíleným jednotlivcem a dojde ke zkáze celého lidstva.

Druhým možným problémem je *„supernerovnost"* mezi příslušníky rodu Homo sapiens. Ta by byla důsledkem masivního kognitivního a fyzického vylepšování člověka – *„enhancementu"*, který by za pomoci genového inženýrství a technologií učinil z části lidí třídu nejen aristokracie, ale přímo bohů. Takto vylepšení nadlidé by neměli trpět žádnými chorobami, stali by se dlouhověkými nebo téměř nesmrtelnými, krásnými, mentálně extrémně pokročilými. Naopak zbytek lidstva by se v porovnání s nimi stal bezvýznamným a zcela postradatelným, obdobně, jako jsou pro dnešního Homo sapiens podřadnými a postradatelnými i naši nejbližší příbuzní, primáti. Můžeme sice namítnout, že rozložení bohatství v dnešním světě již připomíná tuto *„supernerovnost"*, ale díky konečnosti života jednotlivce a meznímu užitku spotřeby nemohou mít lidé stojící na vrcholu společenské pyramidy celkové životní štěstí o tolik vyšší než ostatní. Velké navyšování délky života, schopnosti konzumovat a chápat by tyto limity prolomilo.

Posledním velkým problémem je hrozba, o které v současné době mluví přední světoví veřejní intelektuálové, jako například Bill Gates, Elon Musk nebo Sam Harris. Tato hrozba je známa jako *„superinteligence"*.[6] Tento koncept představuje obecnou umělou inteligenci, která řádově převyšuje lidský intelekt a která je již schopna sama sebe vylepšovat a zvyšovat svůj výkon. *Superinteligence* by tedy v budoucnosti byla zcela superiorní lidské společnosti a bylo by pouze na ní samotné, co by s naším světem a tím pádem s i lidmi provedla, respektive není znám důvod, proč by jí na lidstvu mělo vůbec záležet. Není také zřejmé, jak by ji bylo možné kontrolovat.

Jak tento stručný historický a futuristický výčet souvisí s autonomními vozidly? Společně sdílí motivaci, jež filosofa podněcuje, aby se zabýval společenskými a etickými dopady technických inovací. Někdy totiž až v souvislos-

5 Nejedná se o přírodní zákon, ale o predikci na základě pozorování. A ta má své fyzikální limity.

6 Bezpečností umělé inteligence se zabývají především experti na poli IT, filosofický náhled tohoto problému představil např. Nick Bostrom: Bostrom, N., *Superintelligence: Paths, Dangers, Strategies*. Oxford, Oxford University Press 2014.

ti se *„superproblémem"* vyjde najevo, že je žádoucí si v optimálním případě kvalitativní změny promyslet ještě předtím, než nastanou. Je také třeba důsledně zvážit, kam jako společnost plánujeme směřovat (jaké jsou naše žádosti) a co považujeme za dobré. To všechno je ale nutné udělat ještě předtím, než učiníme nevratné kroky – než sestavíme *„doomsday machine"* do kapsy nebo si postavíme svého vlastního tyrana v podobě *superinteligence*.

Technologické inovace zpravidla předcházejí právním a někdy i morálním normám. I když je tento celkem banální výrok nadužíván (ostatně, proč by naši zákonodárci měli psát zákony týkající se inovací, o kterých ani nevíme, že o nich někdo může uvažovat), nic to nemění na jeho relevantnosti zejména pro AV. Technologie se nachází už ve fázi velmi slibných prototypů nasazených v ostrém provozu, ovšem legislativa v našem domácím prostředí schází a diskuze o morálních aspektech AV ještě nebyla seriózně otevřena. Bude nám stačit jen aplikovat současné morální konvence a etické teorie platné pro řidiče Homo sapiens na „řidiče" – automatické systémy? Soudím, že v případě AV je žádoucí opustit optiku kvantitativních změn a považovat AV jen za jiný příklad řidiče. To nám umožní, abychom řešili daný problém lokálně – z pohledu aktérů–řidičů, kteří budou odstraněni podobně, jako byli odstraněni koně. Zatímco první náhrada za jiný druh pohonu byla jen mechanická, nastávající změna je kognitivní: Nahrazujeme aktéra řidiče něčím jiným. A aplikace toho, jak se dnes řidič chová, snaha jej emulovat, je ještě méně žádoucí, než by bylo přidělávat podkovy na kola svých aut. Cílem této eseje bude proto poukázat na sérii problémů relevantních pro filosofickou diskuzi, které vyvstanou při zavádění nové revoluční technologie autonomních vozidel. Ambicí tedy není problémy vyřešit, ale doporučit témata společenské diskuze, které se kromě filosofů zúčastní výrobci AV, představitelé státní správy a obecních zastupitelstev a také zástupci z řad veřejnosti – uživatelé aut.

Praktickým důvodem této diskuze jsou jistě plány masivního využití této technologie již během našich životů. Nejedná se tedy o futuristický problém, o kterém je možné spekulovat, AV se již staly aktuálním problémem, který je třeba řešit. Dalším důvodem, v tomto případě teoretickým, je to, že nyní bude nutné vyřešit nebo alespoň formalizovat problém praktické etiky zachraňování a odnímání života. Z pouhého hypotetického myšlenkového experimentu, který se krčil v zákoutí *„trolleyologie"*,[7] se stal praktický problém v rámci robotiky a umělé inteligence, na jehož řešení by filosofové měli pro-

7 „*Trolley problem*" se stal součástí nejen úvodních hodin filosofie, ale už i popkultury. Populární zpracování tohoto myšlenkového experimentu a širšího výzkumu podal např.: Edmonds, D., *Would You Kill the Fat Man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton, Princeton University Press 2013.

gramátorům poskytnut nějaký formalizovaný návod, který již nebude mít podobu metafor či proudu nekonzistentních intuicí.

Za nejzásadnější přínos, který by projekt analýzy (formalizace) morálního jednání AV mohl poskytnout naší společnosti, považuji to, že výstupy problémů, které vyřešíme hypoteticky a prakticky v případě autonomních vozidel, pak můžeme analogicky aplikovat do jiných instancí etiky robotiky. Podobně jako je v genetice modelovým organismem octomilka, jejímž prostřednictvím zkoumáme pravidla dědičnosti, funkce genů a jejich interakce, v otázkách praktické etiky robotiky a umělé inteligence by se tímto modelem mohly stát AV. K tomu by bylo třeba splnit následující předpoklady: 1) AV budou nesmírně užitečné; 2) jejich rozšíření bude masivní, čímž se sníží jejich výrobní náklady a AV se stanou ještě dostupnějšími; a 3) interakce AV s předmětem morálky (lidmi) budou těsné a časté. Pokud se nám v budoucnu podaří vyřešit problémy nastíněné v následujících oddílech tohoto textu, budeme nepochybně schopni nabízená řešení aplikovat na další autonomní roboty v nejrůznějších odvětvích, kde dochází k interakci s lidmi a jejich zájmy. Stručně tuto úvahu můžeme shrnout jako řešení problému ve variantě s méně stupni volnosti. Jeden stupeň volnosti „pohybu/regulace" mají dveře, dva výtah a tři robotická sekačka trávníku. To, co vyřešíme jako snadnější problém, následně aplikujeme na problém komplikovanější. Závěry, jež vyplynou ze zkoumání morálních problémů AV, můžeme následně aplikovat na autonomní zbraňové systémy nebo na programování dobré umělé inteligence.

3. Proč jsou AV problémem i pro etika

Mohou být technologie ABS nebo barva vozidla etickým problémem? ABS přebírá od řidiče kontrolu nad jednotlivými koly a při brždění je střídavě zastavuje. K základní podobě tohoto problému nemůže filosof přispět žádnou přidanou hodnotou. Podobně ani zmíněná barva vozu není něčím, k čemu by filosofové mohli říci více než kterýkoliv technik. Můžeme si leda představit, že lépe viditelné barvy jsou bezpečnější, a jsou proto i lepší volbou, pokud chceme minimalizovat riziko neštěstí (zanedbáme-li reflexní prvky). Pasivní systémy nebo systémy minimálně interagující s morálními subjekty a morálními aktéry (lidmi) mohou zkrátka plně existovat i bez diskuze s filosofy. Existuje vůbec v rámci řešení problematiky AV nějaké místo pro filosofa?

„Pokud se něco může pokazit, tak se to pokazí." Samozřejmě, že všechny lidské výtvory mají své limity, a i kdyby zkušení technici učinili svá AV sebedokonalejšími, i malá šance na poruchu dříve nebo později zaviní dopravní nehodu. AV bude mnoho a budou v provozu delší dobu – určitě dojde k dopravní nehodě. V takovém neblahém případě si tvůrce řídicího softwaru přeje, jako

asi každý, aby z daných možných nešťastných scénářů nastal ten nejméně špatný. Jeho zájmem je naprogramovat AV tak, aby před nevyhnutelnou kolizí zvolilo takovou akci, jejíž výsledek bude relativně nejlepší vzhledem k možným alternativám. Uvnitř AV se budou nacházet lidé (cestující), taktéž ve výchozích a cílových bodech budou lidé (kolemjdoucí) – a připomeňme, že jedním z hlavních argumentů obhajujících masové rozšíření AV je snížení počtu úmrtí a zranění způsobených dopravními nehodami. Šance, že k takovému nešťastnému scénáři – kdy má AV např. volbu zabránit větší nehodě obětováním své posádky nebo kdy vybírá mezi dvěma různými kolizemi s fatálními následky pro různé počty chodců – dojde, je možná malá, ale jak již bylo řečeno: i malá šance se vzhledem k pravděpodobné délce doby užívání AV a s ohledem na mnohost aktérů rovná téměř jistotě. Programátoři řídicího softwaru chtějí učinit chování AV „morálním", a tak se obracejí na filosofy, aby zjistili, jaká pravidla rozhodování mají být u AV během nehod použita.

Zastoupení role řidiče může probíhat pomocí emulace. Té bychom docílili sběrem dat z manuálního řízení a kolizních situací. Je ovšem zjevné, že tato varianta má zásadní nedostatek: dopravní kolize totiž řidiči řeší v časové tísni, ve stresu a s velmi omezenou znalostí situace. Taková rozhodnutí budou tedy inkonzistentní a zřejmě nebudou odpovídat ani morálním preferencím řidičů – rozhodnutím, která by učinili po zralé úvaze. Emulace takových nouzových řešení je tedy nežádoucí. Proto je jednou z výhod autonomního řízení skutečnost, že je lze naprogramovat dopředu. Autoři programu mají obrovské množství času na jeho přípravu (rozhodnutí bude deliberativní, nebude pouhým reflexem) a navíc bude postaveno na kolektivním rozhodnutí – vznikne na základě společenské diskuze. Předpisy vytvořené pro rozhodovací software budou vlastně o tom, jak dopadnou dopravní nehody, rozhodovat ještě předtím, než auta vůbec vyrazí. Pasažéři budou moci dopředu vědět, zda jejich auto udělá vše pro jejich bezpečí a zda bude dodržovat dopravní předpisy, nebo zda naopak nebude na posádku brát ohled a v kolizní situaci jen nestranně zváží konsekvence možných reakcí.

4. Řidič se změní z morálního aktéra na pouhý pasivní subjekt

V debatě týkající se etiky AV se pozornost soustředí zejména na typy kolizního systému,[8] které jsou však jen variantou tzv. vozíkového problému.

8 Lin, P., Why Ethics Matters for Autonomous Cars. Gerdes, J. Ch. – Thornthon, S. M., Implementable Ethics for Autonomous Vehicles. Maurer, M. et al., Autonomous Driving: Technical, Legal and Social Aspects. Všechny tři příspěvky jsou součástí sborníku: Maurer, M. – Gerdes, J. Ch. – Lenz, B. – Winner, H. (eds.), *Autonomous Driving*. Berlin, Springer Publishing Company Inc. 2016. Dostupné také online na: https://link.springer.com/book/10.1007%2F978-3-662-48847-8; [cit. 24. 10. 2018].

Autonomní řízení přináší jen „další stupeň volnosti" k v minulosti již obsáhle diskutovanému vozíkovému problému, přičemž je přehlížen jiný eticky relevantní problém. Řidič manuálně řízeného vozu je aktérem (*„moral agent"*), který (doslova) řídí vůz, a může tedy volit – v poli morálního rozhodování – mezi dodržováním silničních předpisů a jejich porušováním. Protože sám sedí za volantem, je odpovědný i za to, co se svým vozem činí. Nemusí se jednat o dramatickou variantu vozíkového dilematu, kdy řidič volí, zda své auto navede ze srázu, nebo najede do velké skupiny lidí, která vyšla z již havarovaného autobusu. Řidiči tradičních vozidel řeší i malá morální dilemata: například zda budou preferovat svůj čas na úkor času druhých (jízda výhradně v levém pruhu, za sanitkou apod.). Vlastník AV ovšem přestane čelit velkým i malým morálním problémům – jeho vůz se bude chovat tak, jako by jeho „tělo" pozřelo morální pilulku, která mu zabrání se chovat nemorálně.[9]

Je možné, že by neochota předat svou morální odpovědnost cizímu programu, navíc posílená nedůvěrou v techniku (jež byla zmíněna ve druhé části eseje), mohla vyústit v rozhodnutí, že si skupina „konzervativních" vlastníků raději ponechá svá manuálně řízená auta, která nebude chtít vyměnit za AV. Proti jejich potenciálnímu rozhodnutí bývá stavěn argument, že přeměna řidiče–aktéra na pasažéra AV je jen analogií toho, co dnes zažívají lidé, když přesednou z vlastního vozu do autobusu městské hromadné dopravy. Tento argument je však poněkud mylný, jak následně ukážu. Jaké rozdíly tedy panují mezi těmito dvěma situacemi? V případě jízdy autobusem je stále přítomen lidský řidič, což může hrát roli u technologicky méně progresivních lidí. Další rozdíl spatřuji v odpovědnosti, která je založena na vlastnictví. Pokud jsme pasažéry autobusu, za jeho chybné jednání, ať už vede ke kolizi nebo k porušení dopravních předpisů, jsou odpovědní řidič a majitel vozu. Pasažéři jsou zcela zproštěni odpovědnosti za další osoby ve voze i za ostatní účastníky dopravního provozu (jsou doslova morálními *„free riders"*). Ovšem v případě, že jsme vlastníky AV a jsme ex-řidiči, stále cítíme zodpovědnost za následky jízdy tohoto vozidla, i když to nejsme my, kdo nakonec „otáčí volantem".

Zavedení AV, bude-li direktivní, učiní ze všech pasažéry a ex-řidiči neponesou žádnou odpovědnost za následky rozhodnutí na silnici; odpovědnost bude přenesena na tvůrce software. Z konsekvencialistické perspektivy to je ovšem jen dobře: sníží se nehodovost a skutečnost, že řidiči přijdou o příležitost zušlechťovat svůj charakter tím, že nepodlehnou pokušení a neporuší

9 Tento problém považuji za typ morálního enhancementu a je vlastně analogický „morální pilulce". Podrobněji k diskuzi o morálním enhancementu viz v: Persson, I. – Savulescu, J., *Unfit for the Future: The Need for Moral Enhancement*. Oxford, Oxford University Press 2012.

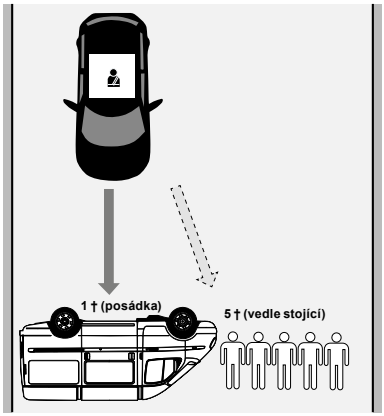
dopravní předpisy (např. objížděním kolony v protisměru), bude vyvážena tisíci zachráněnými oběťmi manuálně řízených aut. Pokud by strach z pozbytí aktérství stále vyvolával neochotu vyměnit své tradiční auto za AV, nabízím dva návrhy. První řeší odpovědnost za vlastnictví vozu. Řešením by byl *„car sharing"* – ex-řidič by si v „půjčovně" objednal na svou cestu auto, které by jej v místě A vyzvedlo a dovezlo jej na místo B. Takový pasažér by nebyl vlastníkem vozu, a tak by necítil odpovědnost za činnost pronajatého vozidla. (Podobně jako když dnes cítíme, že je poměrně velký rozdíl mezi tím, zda autem srazíme srnu my sami, nebo se to přihodí řidiči autobusu, ve kterém sedíme.) Druhým způsobem řešení je nabídnout více druhů „morálních pilulek" – čili více typů obecného chování autonomních aut. (Tím není myšleno, že by za příplatek AV umělo na požádání vybrždďovat nebo svítit protijedoucím řidičům do očí dálkovými světly.) Touha po volbě, po aktérství (sice nepřímém) by mohla být naplněna výběrem z několika typů „morálních řídicích programů", o kterých bude řeč v následující kapitole.

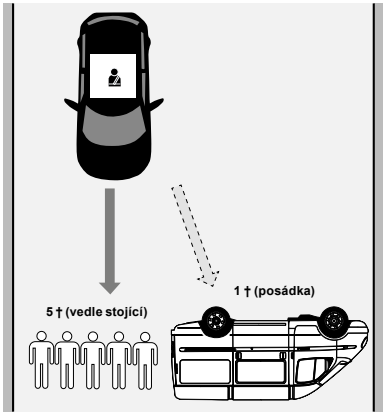
5. Myšlenkové experimenty, metodická redukce

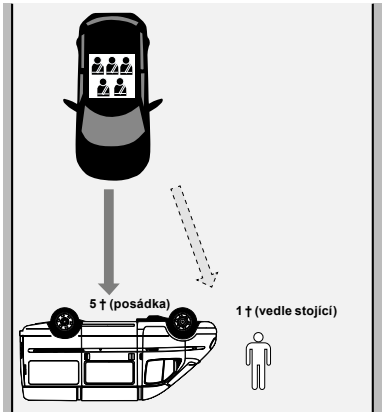
Pro potřeby následující argumentace si definujme tři zjednodušené typy morálního softwaru. Jsou to ilustrativní příklady, pomocí nichž si vysvětlíme, jaké rozdíly v nastavení softwaru mohou hrát roli při jeho preferenci mezi cestujícími (nebo třeba vládami). Představme si hypotetickou dopravní nehodu v tunelu. Černé autonomní auto s vyznačeným počtem posádky má poruchu brzd a může volit jen ze dvou variant: první je neměnit směr jízdy a jet rovně, druhou variantou je zahnout. Následky jsou jediné dva možné: černé auto po nárazu do prázdné dodávky havaruje a jeho posádka (černí panáčci) zahyne. Druhou variantou je, že černé AV zahne a „zabrzdí" nárazem do osob v tunelu (bílých panáčků), čímž je sice zabije, ale jeho posádka–černí panáčci vyvázne bez újmy. Tato modelová situace nemá emulovat realitu a nabízené typy softwaru nemusí nutně reprezentovat konkrétní etické teorie. Obdobně jako v mechanice někdy např. redukujeme tělesa do hmotných bodů a zanedbáváme tření vzduchu, můžeme v morální filosofii zjednodušit složitý svět do myšlenkového experimentu jen s jednou binární volbou.

Nyní zavedeme tři typy řídicího softwaru: *Tank,* který vždy preferuje životy posádky AV před životy kolemjdoucích a pasažérů jiných aut. *Počítadlo*, které vždy preferuje takové řešení krizové situace, při němž zemře nejméně lidí, respektive nejvíce je zachráněno. Posledním je *rytíř*, který nikdy aktivně nezmění směr jízdy, pokud by tato akce měla za následek zabití lidí, kteří by při absenci akce nebyli nijak ohroženi. Z nákresů je patrné, že jízda bez změny směru, vedoucí k úmrtí posádky nebo chodců, je ponecháním zemřít (*„letting die"*), zatímco jízda se změnou směru má za následek usmrcení

(*„killing“*) – které sice není zamýšlené, ale je předvídatelné. (V reálných situacích si představuji např. takovou verzi počítadla, která sníží bezpečnost vlastní posádky o 5%, aby o 60% snížila riziko hromadné dopravní nehody, při níž by zahynulo deset lidí. Předpokládám, že reálné programy budou mít podobu smíšených strategií kombinujících všechny tři uvedené typy softwaru.)







Na prvním obrázku by *počítadlo* a *rytíř* nezměnili směr jízdy, *tank* by zabočil vlevo. Na druhém obrázku by *tank* a *rytíř* nezměnili směr jízdy, *počítadlo* by zabočilo vlevo. Na třetím obrázku by *rytíř* nezměnil směr jízdy, *počítadlo* a *tank* by zabočili vlevo. Z těchto předpokladů budeme dále vycházet.

Ilustrace jsou mé vlastní. Při jejich tvorbě jsem použil obměněné obrázky bez autorských práv (pixabay.com) a program pro ne-profesionály MS Paint.

Etikům zadali inženýři následující úkol: Je nutné naprogramovat řídicí algoritmus a přejeme si, aby auta byla „morální". To je problémem a výzvou zároveň. Problémem je to zejména proto, že mezi profesionálními filosofy ani jinými mysliteli nepanuje obecná shoda na jediné správné teorii, na základě které by bylo možné takový software napsat. Výzva je to v tom smyslu, že máme naše morální soudy zpracovat do programu, čili máme odstranit veškerá subjektivní východiska a zároveň připustit pluralitu názorů. Tato výzva chce pro všechny definovat jeden závazný systém, který by zpracovával informace z veřejných, sdílených dat senzorů. Vnímám to jako příležitost k tomu, aby společnost znovu promyslela základy morálky. Část etiky se bude muset konečně transformovat ze subjektivního pohledu na operacionalizovatelnou teorii zapsatelnou do objektivních matematických vzorců, a to z čistě praktických důvodů: do palubního počítače bude rozhodování třeba zapsat ve formě jedniček a nul. Tato výzva může konvenovat konsekvencialistům zaměřeným na kvantitativní měření potěšení. Jiným etickým školám, zvláště těm, které operují s hodnotami ctností a povinností, by to ovšem mohlo činit problémy, a to až do té míry, že by takto vynucený kalkulus nemusely vůbec považovat za eticky vhodný.

Inženýři se obrací také na experimentální vědce (zejména morální psychology). Žádají je, aby popsali, jak probíhá morální soud lidí v případech volby softwaru – ať už při hypotetickém nákupu AV nebo při hypotetickém tvoření legislativy. Osobně se proto domnívám, že je třeba se více věnovat morálním intuicím. Netvrdím, že si všichni softwaroví inženýři uvědomují, že je morální soud zkoumaný psychology deskriptivní, zatímco etika, které se věnují filosofové, je preskriptivní. Ve třetí a čtvrté části této eseje jsou ostatně uvedeny důvody, proč by emulace lidského soudu byla kontraproduktivní. Nástrojem zkoumání morálních intuicí ve vozíkovém problému jsou tradičně myšlenkové experimenty spojené s dotazníkovým šetřením, dalšími metodami může například být užití virtuální reality[10] nebo magnetické rezonance.[11] Smyslem těchto výzkumů je poodhalit faktory, které se podílejí na tvorbě morálního soudu, a dále zjistit, jaká kognitivní zkreslení ovlivňují naše rozhodnutí. Pomocí takto získaných výsledků budou automobilky a vlády v budoucnosti moci např. predikovat to, jaký typ morálního softwaru bude nejvíce žádán na trhu, připustí-li volbu. Zastoupení různých typů morálního jednání AV není však ve společnosti jen jednosměrný proces. Noví vlastníci se totiž budou moci rozhodovat i na základě toho, jaká auta ve společnosti převládnou:

10 Navarrete, C. D. et al., Virtual Morality: Emotion and Action in a Simulated Three-dimensional "Trolley Problem". *Emotion*, No. 12. 2. 2012, s. 364.

11 Greene, J. D. et al., An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293, 2001, No. 5537, s. 2105–2108.

Pokud bude normou sobecké vozidlo typu *tank*, bude to mít na volném trhu AV pravděpodobně za následek „závody ve zbrojení", v nichž půjde o co nejbezpečnější a nejmohutnější *tanky*. Pokud naopak ve společnosti budou od začátku převládat altruistická auta typu *počítadlo* a budou zavedeny sankce pro vlastníky sobeckých vozidel (ať už daňové nebo jen společenské v rámci snížení prestiže), je šance, že dojde k naplnění konsekvencialistického optima a při dopravních nehodách začne umírat méně lidí.[12]

6. Praktické problémy AV

Jedním ze způsobů volby morálního softwaru AV je odmítnutí celého tohoto systému, ve kterém je dopředu rozhodnuto o tom, kdo v kolizní situaci pravděpodobně přežije (díky preferenci vozu) a kdo naopak zemře. Únikem z potenciální pasti tohoto problému by mohlo být setrvání u manuálně řízeného vozidla. Taková úvaha je však založena na chybném předpokladu, že jsou si tradiční auta rovna. Ve skutečnosti však již dopravní prostředky typu *tanku* i *počítadla* existují. Vezmeme-li v úvahu jen bezpečnost posádky, má při kolizích největší výhodu řidič mohutného vozu s vysoko posazeným sezením (nákladní vůz). O něco méně bezpečná jsou robustní SUV, ještě méně bezpečné jsou vozy nižší třídy a vůbec nejhůře na tom jsou při kolizích řidiči motocyklů. Faktem tedy je, že už dnes při pořizování vozů kalkulujeme s naší bezpečností. A nekalkulují jenom zákazníci automobilek. Tradičním učebnicovým příkladem kalkulací s lidskými životy byl vůz Ford Pinto. Automobilka Ford vyčíslila, že by náklady na zesílení přepážky kolem palivové nádrže u všech vozidel modelu Pinto byly vyšší než kompenzace za několik zemřelých, kteří uhoří po jejím vznícení. Také vlády zemí kalkulují s lidskými životy v dopravních prostředcích: Letadla jsou sestřelována, pokud ohrožují například elektrárnu nebo směřují na jiný strategický cíl. Otázka tedy zůstává: Jaký typ softwaru pro AV si pořídit?

Z nedávného amerického výzkumu jednoznačně vyplývá,[13] že by participanti preferovali, aby ostatní účastníci silničního provozu měli auta typu *počítadlo,* ale pro sebe by si vzali typ *tank*. Zároveň by nechtěli, aby stát všem nařizoval typ *počítadlo*. Tato volba – chtít po všech ostatních, aby byli altruističtí a sám zůstat sobecký – se může zdát ekonomicky racionální. Analogicky tomuto přístupu odpovídá přístup jednotlivců k americké Národní bezpečnostní agentuře (NSA). Jednotlivci nevadí, že NSA sbírá telefonní data mnoha

12 Ekonomické hry na „common good" ukazují, že pokud se začnou objevovat sobecké strategie, altruistické jednání se udrží na jen základě trestání. Fehr, E. – Gächter, S., Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90, 2000, No. 4, s. 980–994.

13 Bonnefon, J.-F. – Shariff, A. – Rahwan, I., The Social Dilemma of Autonomous Vehicles. *Science*, 352, 2016, No. 6293, s. 1573–1576.

dalších lidí. Člověku většinou začne vadit až to, že NSA sbírá právě jeho data. (Na jednu ze souvisejících obav o soukromí, totiž že díky AV budou cestující sledováni, lze odpovědět tak, že sledování AV se moc neliší od kontroly pohybu mobilního telefonu, která je dnes již běžná.)

Dalším praktickým problémem je rozhodnutí, kdo má být zodpovědný za tvorbu softwarových programů morálního rozhodování AV. Stát musí dopředu vytvořit legislativu provozu AV, protože AV se brzy zařadí do běžného provozu a stanou se účastníky dopravních nehod. V případě právního vakua by zúčastněné strany zřejmě žalovaly úplně všechny. Pokud přenecháme rozhodování o typech softwaru jen automobilkám, nemáme jistotu, že software bude řídit AV stejně, jak to prezentují. Vzpomeňme na nedávnou *„Dieselgate"* německé automobilky Volkswagen. Víme-li, že automobilky podvádějí s výfukovými plyny, nemůžeme tušit, zda nebudou podvádět i s morálními algoritmy. Co když budou auta Volkswagenu v kolizních situacích vždy preferovat záchranu svých pasažérů oproti jiným zúčastněným?

Posledním z praktických problémů AV je mediální zkratka *„Build to kill"*, tedy přisuzování AV při kolizních situacích záměr přejíždět kolemjdoucí. Ovšem i jiný výtvor inženýrů – most – je *„Build to kill"*, pokud bude přetížen nad limit nosnosti. O mostu bychom ovšem neřekli, že zabíjí, ale že má technické limity. (Pokud si nepřejeme most s určitou nosností, můžeme vodní plochu protnout valem ze zeminy. Protože cílem inženýra není vybudovat ideální dílo pro všechny situace, ale vytvořit věc určenou k danému účelu a v daném rozpočtu, čili v limitech užití.) Obdobně rozdíl mezi úmrtím při běžné dopravní nehodě a „zabitím autonomním vozidlem" bude asi přisuzován intenci zabíjet, která bude vnímána jako atribut autonomního vozu. Analogicky je tomu například i s konzumací masa, která je běžně tolerována. Avšak pokud se americký zubař vydá do Afriky zastřílet si pro zábavu na lvy, což obsahuje intenci zabití (nikoliv jako prostředek), už je považován za morálně zkaženého člověka. Můžeme cítit spravedlivý hněv, když zjistíme, že řidič úmyslně někoho přejel, aby se vyhnul větší nehodě, ale AV se při přejetí člověka, k němuž dojde za účelem záchrany jiných lidí, mnohem více podobá mostu s limitním zatížením.

7. Závěr

Kvalitativní změna, která nastala výměnou vozu taženého koňmi za automobil, nachází svou obdobu v dnešním přechodu k AV jen velmi vzdáleně. Neměníme řidiče z masa a kostí za jiného řidiče z mikročipů, ale rozhodujeme morální problémy mnohem dříve, než k samotné nehodě vůbec dojde. Další ještě zásadnější změnou bude to, že řidiči budou zbaveni přímé morální odpovědnosti, jakož i možnosti morálně se rozhodovat v dopravních situacích.

Bude třeba nasbírat také data týkající se morálního soudu při morální a skutečné preferenci (vlastním nákupu autonomního vozidla).

Navrhuji zaměřit se na právní jistotu v oblasti kontroly morálního softwaru AV. K tomu je nutné co nejrychleji zjistit, „co chceme přijmout jako normy" řídící provoz AV, a připravit ještě v předstihu příslušné zákony. Z provedených empirických studií zatím vyplývá, že čistý konsekvencialismus bude nutné doplnit vyváženou preferencí posádky. Pokud by stát striktně trval na typu *počítadlo,* bylo by možné, že by taková AV nevytlačila ze silnic nebezpečné tradiční vozy. Tento model a smíšené strategie z části *počítadla* a z části *tanku* by mohly zabránit tomu, aby se fixoval suboptimální stav úplné převahy vozů typu *tank*.

V této eseji byly záměrně vynechány problémy týkající se dalších specifikací morálního softwaru, které jsou ale oproti zmíněnému hlavnímu problému podružné. Jedním z nich je dodržování silničních předpisů. Pro vůz typu *počítadlo* by se stala problematickou kupříkladu volba mezi kolizí s motorkářem s přilbou a motorkářem bez přilby, neboť by preferoval kolizi s prvně jmenovaným, kterému by sice pravděpodobně způsobil menší škodu, ovšem „trestal" by jej takto za dodržování předpisů. Smyslem výše uvedených modelových příkladů softwaru nebylo hledání globálního řešení, ale poskytnutí definic a příkladů k diskuzi. Druhým opomenutým problémem bylo využití standardizovaných panáčků, neboť skutečné kalkulace budou zřejmě počítat i zachráněné roky života a jejich potenciální kvalitu.

Závěrem shrnuji, že na základě konsekvencialistických východisek není nutné přijmout typ *počítadlo*, protože hlavní morální problém leží v současnosti jinde – skutečným dilematem je otázka, kdy se konečně zbavíme manuálně řízených aut, abychom zachránili tisíce životů. Dokonce i ne-konsekvencialistické řešení dané situace, tj. preference jakýchkoliv typů morálního softwaru – i toho „bez iniciativy" zvaného *rytíř* –, by vedla k záchraně tisíců životů, jejichž ztráty jsou dnes zapříčiněny chybami řidičů. Mohli bychom dosáhnout snížení počtu obětí nehod až o 95 %. Musíme brát v úvahu i to, že po zavedení jakéhokoliv morálního softwaru do AV lze jeho program následně vylepšovat. Stačí softwarová aktualizace a z *tanku* můžeme vytvořit *počítadlo* nebo případně jakoukoliv kombinaci rozhodování. Po otevření veřejné diskuze bude nutné se zaměřit i na další problémy. V jejich hierarchii, hned za hlavním úkolem – zbavit se manuálně řízených aut –, vyvstává druhý nejdůležitější: Jak se vyrovnat s obavou ze ztráty morálního aktérství.