důsledky nedohlédne, snad s výjimkou krádeže identity, což je sám o sobě abstraktní pojem zahrnující celou řadu podvodů.

Nemá smysl se domnívat, že veřejnost by měl současný stav dat přivádět k zuřivosti; i když bychom se před zneužitím dat měli mít na pozoru vždy, spravedlivá a konsensuální výměna údajů umožní, aby se cenné technologie dostaly do rukou miliard lidí. Technologové by si jen naběhli, kdyby vyvolávali odpor proti všem inovacím závisejícím na datech. Udělají lépe, budou-li se zabývat tím, proč je veřejnost vůči datům lhostejná.

K hlavním příčinám této lhostejnosti patří nedostatek porozumění; nejúčinněji veřejnost zapojíme tím, že podpoříme její znalosti v oblasti dat a její sebedůvěru. Pomůže nám jak zhmotnění dat, tak nabízení technologií na podporu ochrany soukromí (PET) a vysvětlování jejich hodnoty. Měli bychom se však také pokusit zvýšit informovanost veřejnosti prostřednictvím vzdělávání a zpravodajství. Technologové musí upozorňovat i na jiné problémy s daty než na problémy se zabezpečením; mohou nabídnout více než jen obrázky visacích zámků. Měli bychom se snažit vytvořit přesvědčivý veřejný a politický narativ o roli dat v naší budoucí společnosti a zabývat se jejich rostoucím významem a hodnotou („Dostáváte za ně spravedlivou kompenzaci?"), riziky, jež předpokládané a agregované údaje mohou vytvářet, a hrozbou algoritmické nespravedlnosti způsobené zneužíváním údajů.

Toho lze dosáhnout tím, že využijeme politický proces. Zainteresovaní technologové mohou spoluutvářet státní datovou politiku tím, že se budou podílet na konzultacích, psát voleným zástupcům, či dokonce kandidovat. Ti, kteří k politice tolik netíhnou, mohou vytvářet řadu technologií. Technologie na podporu ochrany soukromí (PET), například blokátory reklam (jsou-li etické!), anonymizéry, jednorázové účty, šifrovací software a nástroje pro automatizaci žádostí podle zákona o ochraně údajů, zvýší informovanost veřejnosti a její odolnost. Vždy budou zapotřebí také alternativy ke službám pro získávání údajů, jež budou sloužit jako zářné příklady toho, jak lze s údaji a se soukromím zacházet spravedlivě.

5

Pohled novýma očima

Donedávna jsme předpokládali, že mají-li stroje porozumět fyzickému světu, svět jim bude muset dobrovolně poskytnout informace. K tomu bychom potřebovali lidské kartografy – informační architekty a další mistry taxonomie a označování – i celou řadu automaticky se popisujících a hlásících se objektů („spimes"), jež by neustále vysílaly informace o svém stavu.

Zdá se však, že to již není nezbytné. Jak naznačuje teorie zprostředkování, každá inovace přináší nové způsoby interpretace světa. Rodící se technologie jsou čím dál schopnější získávat informace z fyzického světa jednostranně tím, že ze stínů vytahují to, co bylo dříve neviditelné.

Počítačové vidění

Nejdůležitějším vstupním zařízením příštího desetiletí budou kamery: ještě levnější a menší, namontované v osobních zařízeních, v našich obývacích pokojích, na dronech i v našich ulicích. Podle odhadů jedné firmy jejich počet do roku 2022 dosáhne 45 miliard, což představuje nárůst o 220 % za pouhých pět let. Za peníze si dnes koupíte denní satelitní záběry odkudkoli na světě, nositelné fotoaparáty budou brzy přenášet miliony tváří

a autonomní vozidla budou zaznamenávat města ze všech možných úhlů. Brzy budeme mít tolik pohledů na téměř jakoukoli akci, že bude důležité vědět, kde kamery nejsou – pravděpodobně půjde o místa, kde se budou skrývat nejtemnější tajemství společnosti: jatka, skládky a zasedací místnosti mocných.

Počítačové vidění, které tyto záznamy převádí do strukturovaných dat, se může osvědčit jako „killer" aplikace informačního věku. Rozpoznat čárové kódy, text a čísla umí stroje už řadu let; teď otevřou oči dokořán.

Rozpoznávání obličejů v těch správných souvislostech je již pozoruhodně přesné. Systém DeepFace společnosti Facebook má při rozhodování o tom, zda je na dvou fotografiích jedna a táž osoba, 97% přesnost a podle řady poskytovatelů téměř dosahuje lidských schopností. Několik start-upů v oblasti rozpoznávání obličejů má hodnotu více než jedné miliardy dolarů; některé, například Megvii, mají privilegovaný přístup k datovému souboru 1,3 miliardy tváří, jenž patří čínské vládě. Stále je však třeba překonat závažné výzvy. Výzkumnice Joy Buolamwiniová z Massachusettského technologického institutu (MIT) zjistila, že řada nástrojů pro rozpoznávání obličejů má zkreslené výsledky a špatně identifikuje 1 % mužů světlejší barvy pleti a 35 % žen tmavší barvy pleti[104] a že přesnost nabízená zasvěcenci platí pouze za určitých podmínek. Facebookové algoritmy pro označování fotografií mají například výhodu v podobě ohromných trénovacích souborů dat a předchozí znalosti sítí přátel; systémy pro rozpoznávání obličejů prodávané kvůli zabezpečení zase často vyžadují, aby se uživatelé dívali přímo do kamery. Zatím nelze rozpoznávat obličeje v rámci celého města; zpracování a ukládání dat vyžaduje příliš velkou paměť a starší průmyslové kamery mají příliš nízké rozlišení. Možnosti rozpoznávání se však časem budou jen zlepšovat.

Řada technologických firem, jež v počítačovém vidění spatřuje zlatý důl, se zoufale snaží přesvědčit uživatele, aby se zpracováním obličeje souhlasili. Nařízení GDPR považuje biometrii za údaje „zvláštní kategorie", což znamená, že uživatelé k uchovávání a zpracování těchto údajů musí dát svobodně výslovný souhlas, s výjimkou několika případů ve veřejném zájmu, například policejního dohledu. Technologické firmy proto musí uživatelům

[104] Steve Lohr, „Facial Recognition Is Accurate, if You're a White Guy", *The New York Times*, 9. února 2018, nytimes.com.

nabídnout zvláště lákavé výhody; naštěstí pro ně počítačové vidění přináší bezpočet možností inovativního využití.

Počítačové vidění může samozřejmě klasifikovat nejen tváře, ale i celá prostředí. Lidar – v podstatě radar využívající světelné impulzy – strojům umožňuje mapovat okolí a je velmi důležitý pro většinu systémů autonomních vozů. Postupem času budou miliardy umělých očí vytvářet minimální ohraničující pole kolem dopravního značení, popelníků i racků. Dnes si za 249 dolarů koupíte „první videokameru pro vývojáře na světě, která podporuje hluboké učení" – vyžaduje sice montáž, ale dopravu máte zdarma. Odpověď na nejčastější otázku zákazníků „Odhalí i exhibicionisty?" zní prostě ano.

Naslouchající stroje

Ze strojů se také stávají zdatní posluchači. Zařízení Audrey („automatický rozpoznávač číslic") v roce 1952 dokázalo rozpoznat číslice vyslovené majitelem, pro skutečné zvýšení přesnosti však musely přijít Markovovy předpovědní modely a zpracování údajů v cloudu. Dnes, kdy má jakékoli zařízení přístup k výkonnému zpracování a dokáže provést kvalifikovaný odhad toho, co může následovat, je přesnost hlavních platforem rozpoznávání řeči vyšší než 95 %.

Teoreticky je řeč skvělým způsobem komunikace se stroji. Je rychlá – asi sto padesát slov za minutu oproti průměrným čtyřiceti slovům při psaní –, osobní a vhodná pro překonání zmenšující se propasti mezi lidmi a technikou. Má však jistá omezení. Hlasová rozhraní uživatele nutí, aby si pamatovali příkazy, problém však představuje nejednoznačnost i přízvuk a vody dále kalí také intonace.

Rozpoznávání zvuku zvuk kupodivu nepotřebuje. Algoritmus odezírání ze rtů vytvořený vědci z Oxfordské univerzity a ze společnosti DeepMind ve zpravodajských klipech BBC správně rozpoznal 46,8 % slov: byl tedy daleko přesnější než profesionální odečítač ze rtů.[105] Vědci zjistili, že citlivé gyroskopy chytrých telefonů reagují na vnější hluk, a umožňují tak

[105] Joon Son Chung et al, „Lip Reading Sentences in the Wild", *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017.

odposlouchávat okolní rozhovory, aniž by bylo nutné žádat oprávnění k ovládání mikrofonu od operačního systému,[106] a studie Massachusettského technologického institutu (MIT) z roku 2014 rekonstruovala píseň zahranou ve zvukotěsné místnosti pomocí analýzy vysokorychlostního videa rostliny v květináči, která s sebou trhala v rytmu hudby.[107]

Není tedy divu, že naslouchání počítače dává vzniknout úzkostlivým konspiračním teoriím. Zvláště vytrvalé jsou zvěsti, že Facebook slídí prostřednictvím mikrofonů v našich telefonech a poslouchá naše nejtemnější tajemství, aby mohl lépe vytvářet reklamy na míru. Přestože Facebook popírá, že by něco takového dělal, a nejsou pro to ani důkazy, jde o pochopitelnou neurózu. V roce 2017 vědci věnující se oblasti zabezpečení zveřejnili zjištění, že 234 aplikací systému Android „bez vědomí uživatele neustále poslouchá ultrazvukové vysílače v pozadí,[108] a společnost Samsung před dvěma roky připustila, že její chytré televizory sdílely zvuk s nejmenovaných třetími stranami. Výrobci bryskně odpovídají, že budou zvuk zpracovávat pouze na vyžádání – obvykle po spouštěcím slově, například „Alexo" nebo „Ahoj Siri", nebo po stisknutí tlačítka a vždy budou poskytovat zpětnou vazbu, například ikonu mikrofonu či rozsvícenou diodu LED. V neviditelném světě výměny údajů se však tato tvrzení ověřují těžko. Uživatelé musí zkrátka výrobcům slepě věřit.

Po omezení chyb a kolonizaci většího počtu domácností inteligentními reproduktory nás v budoucnosti čeká rozpoznávání, jež bude všude dostupné a bezproblémové. Z naslouchajících asistentů, kteří již nebudou vázáni na domácí zařízení či sluchátka, se stanou všudypřítomné služby. Jakmile technologie překonají takzvaný fenomén koktejlové party a naučí se identifikovat více hlasů, můžeme si představit asistenty, kteří brzy uniknou ze svých kouzelných lamp a stanou se avatary bez těl, jež nás budou doprovázet z našich domovů do veřejné sféry: na ulici, do kanceláře, do baru.

[106] Yan Michalevsky et al, „Gyrophone: recognizing speech from gyroscope signals", zápis z 23. konference USENIX, Security Symposium (SEC'14), 2014.

[107] Abe Davis et al, „The Visual Microphone: Passive Recovery of Sound from Video", *ACM Transactions on Graphics* (Proc. SIGGRAPH), 33, 4, 2014.

[108] Daniel Arp et al, „Privacy Threats through Ultrasonic Side Channels on Mobile Devices", *IEEE European Symposium on Security and Privacy* (EuroS&P), Paříž, 2017.

Rozhovory se stroji

Jestliže stroje naslouchají, mohli bychom s nimi navázat rozhovor. Strojová řeč vygenerovaná laboratoří je dnes stěží rozeznatelná od té lidské a konverzační stroje mohou počítač přimět, aby na požádání šeptal, odmlčel se či změnil výšku tónu. Kromě důsledků pro informační důvěryhodnost s sebou nese umělá řeč také nepříjemné otázky týkající se etiky antropomorfismu. K těm se dostaneme později.

Strojová konverzace může mít na člověka zvláštní dopady. Jestliže s hlasovým asistentem strávíte nějaký čas, zjistíte, že strojené fráze fungují často lépe než celé věty, a rychle se naučíte nastavovat budík na určitou hodinu a devětačtyřicet minut, nikoli na „padesát", protože v angličtině si asistent toto slovo snadno splete s číslovkou „patnáct".

> Lidé mluví se stroji jinak než s lidmi. Přecházejí na jiný rejstřík. Stojíte-li vedle někoho na letišti nebo na autobusové zastávce či podobném místě, obvykle poznáte, kdy mluví spíše se strojem než s člověkem. – Alan Black[109]

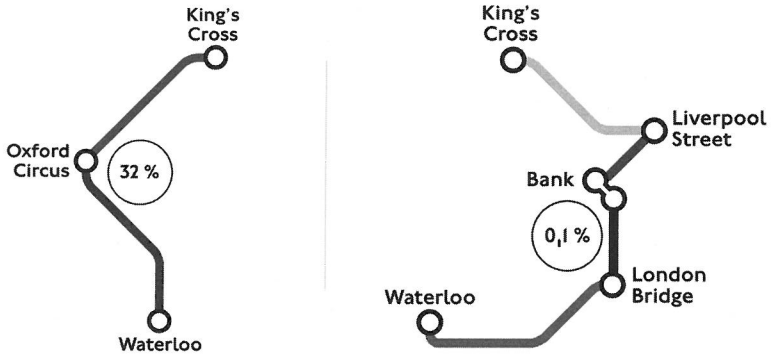
Není žádnou novinkou, že lidé přizpůsobují chování technice. Celá desetiletí gestikulujeme na světelná čidla v koupelnách a vrčíme na nabídku call center a pro optimalizaci vyhledávačů jsme vytvořili celý profesionální obor, který se snaží, aby lidská komunikace byla pro algoritmus přitažlivější. Verbální optimalizace by se však mohla rozšířit i na komunikaci člověka s člověkem, zejména pokud ji zprostředkovává technologie. Řada aplikací pro zasílání zpráv a e-mailů dnes nabízí přednastavené prediktivní odpovědi: „To zní dobře", „Díky za informace!" Ty se rámují jako vhodné návrhy a uživatel má nad psaním vlastních odpovědí plnou kontrolu. Snadno si však představíme systém na základě hlasu, který bude omezený médiem a bude nabízet jen tyto předpřipravené odpovědi. Bude to bezpochyby efektivní, ovšem komunikaci to může ochudit. Tyto konverzační systémy se budou pravděpodobně učit na velkých korpusech textů a odpovědí jiných uživatelů, a budou-li tato tréninková data jazykově a dialektově rozmanitá, může se

[109] Tom Dart, „Y'all have a Texas accent? Siri (and the world) might be slowly killing it", *The Guardian*, 10. února 2016, theguardian.com.

stát, že nestandardní věty z naší mluvy zmizí. Časem se naučíme vynechávat z příkazů hovorové výrazy a stroje budou zase reagovat ve standardizovaných dialektech. Až se sami naučíme komunikovat strojově čitelnými způsoby, v angličtině se můžeme stát svědky toho, že na základě jazyka metropolitní oblasti San Francisco Bay vznikne lingua franca, jež bude pomalu ochuzovat jazykovou rozmanitost.

Datafikované tělo

K mání jsou i údaje o našich tělech. Váhy připojené k síti, inteligentní zubní kartáčky a matrace monitorující spánek tiše dokumentují hříchy těla, fitness náramky měří naše vitální funkce, například tepovou frekvenci a obsah kyslíku v krvi, a naše osobní zařízení stále více odhalují naši polohu. Patent společnosti Facebook popisuje využívání telefonních akcelerometrů a gyroskopů k výpočtu nejen toho, kdy spolu někde stáli dva lidé, ale i toho, kdo stál komu tváří v tvář.[110] V jednom experimentu roku 2016 sledovala společnost Transport for London chytré telefony podle adres MAC, aby porozuměla jejich pohybu po příměstských trasách londýnského metra, a odhalila přitom informace o trase, které se nepodařilo zjistit prostřednictvím systému placeného parkování. Asi 0,1 % cestujících, patrně pod vlivem zlých



[110] Kashmir Hill and Surya Mattu, „Facebook Knows How to Track You Using the Dust on Your Camera Lens", *Gizmodo*, 11. ledna 2018, gizmodo.com.

duchů či halucinogenních psychotropních látek projíždělo mezi stanicemi King's Cross a Waterloo linkou metra Central Line.[111]

Ačkoli se tyto tělesné datové body zdají samostatně bezvýznamné, i triviální informace mohou v kombinaci s jinými soubory dat a při dlouhodobém sledování vypovídat o mnohém. A znovu platí, že zařízení zřejmě nebudeme potřebovat dlouho. Brzy nám ke sledování něčí polohy postačí flotila kamer pro rozpoznávání obličejů, která navíc zjistí, s kým daný člověk zašel na kávu. Již dnes můžeme ze záběrů ve vysokém rozlišení zjistit něčí pulz[112] a získat vzor jeho otisků prstů.[113] Zatímco rozlišení se zvyšuje a na spotřebitelský trh se vkrádá specializovaný hardware – základní termokamery dnes stojí necelých 200 dolarů –, kamery se čím dál častěji dívají nejen na nás, ale i do nás.

Naše těla jsou také nositeli emocionálních informací. Odpudivý potenciál neuromarketingu – využití skenování mozku a kožních sond k měření emoční reakce člověka při pohledu na určitou značku – se v kombinaci s algoritmickým získáváním dat stávají ještě neodbytnějšími. Desítky technologických firem tvrdí, že dokážou kvantifikovat emoci z textu, obrázků, videa a řeči, a tyto technologie se již hojně využívají v herním, reklamním a bezpečnostním průmyslu. Společnost Facebook Australia podle uniklého dokumentu z roku 2017 dokáže zjistit, kdy se mladý uživatel cítí „bezcenně", „hloupě" nebo „nejistě". PR tým společnosti tuto technologii hájil s tím, že k cílení reklam nikdy sloužit neměla.

Hypermapa

Lze si představit, že brzy dokážeme kvantifikovat celý svět a lidi, objekty a vlastnosti proměníme ve strukturované užitečné informace. Nemusíme přitom na každý fyzický objekt lepit QR kódy nebo nálepky RFID, počítačové vidění a naslouchání spolu s obrovskými přívaly dat z různých zařízení zmapují svět na dálku a násilím. V plně poznaném, kodifikovaném a označeném světě budeme přesně vědět, co věci jsou a kde přesně se nacházejí. Můžeme si dokonce

[111] Transport for London, „Review of the TfL WiFi pilot".

[112] Hao-Yu Wu et al, „Eulerian video magnification for revealing subtle changes in the world", *ACM Trans. Graph.* 31, 4, čl. 65 (červenec 2012).

[113] Chris Wood, „WhatsApp photo drug dealer caught by ‚groundbreaking' work", *BBC News*, 15. dubna 2018, bbc.co.uk.

představit jakési matrixové skákání mezi dvěma pohledy: mezi atomy a informacemi. Toto prostředí celkových metadat nazvěme *hypermapou.*

Hyperzmapovaný svět je nabitý potenciálem. Už nikdy se v něm nic neztratí, automobilovými díly počínaje a klíčky od auta a zloději aut konče. Naše tělo se stane nástrojem naší identifikace u dveří do bytu, u odletové brány i při převodu peněz. Kamery budou sledovat únavu řidičů a poruchovost montážních linek a nemoc rozpoznají dokonce dříve, než pocítíme příznaky. Neznámá místa budou rázem známější, překrytá překlady a kontextem pro cestovatele. Hypermapa se přitom nemusí omezovat jen na současnost; díky zkoumání historických snímků, videa a zvuku mohou počítače strukturovat i minulost. Hypermapa rozprostřená v čase by mohla předefinovat naše dějiny způsobem, o jakém se nám ani nesnilo.

Možnosti hypermapy jsou silně ambivalentní a měly by vyvolávat vzrušení i strach. Každá její světlá stránka by mohla být v jiném kontextu škodlivá: ideální technologie pro chytání zločinců je ideální také ke sledování a svět dokonalých informací je svět zralý pro totalitu. Spojíme-li hypermapu s výkonnými technologiemi přesvědčování, můžeme si představit propojené technologie, jež nebezpečně i banálně vzdorují politickým odpůrcům. Vaše zařízení samozřejmě oznámí každé vaše slovo policii, trestem za úchylné myšlenky vám bude ztrojnásobení ceny letenek, a topinkovač vám dokonce spálí snídani.

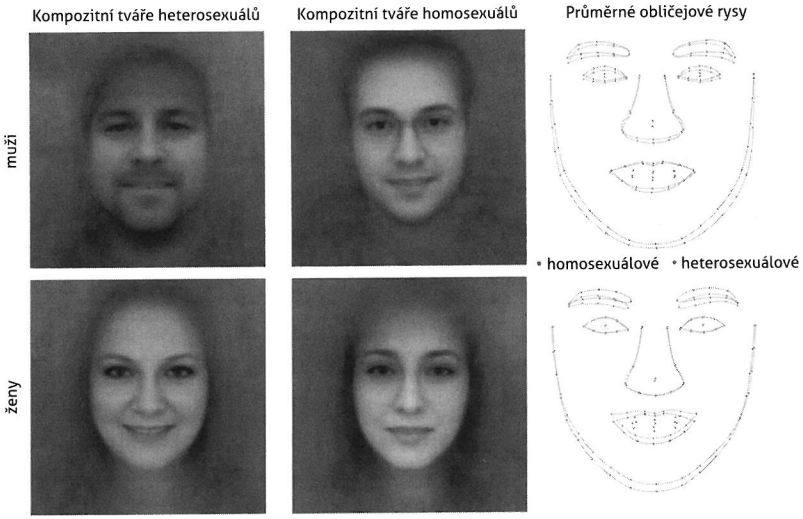
Samotná hypermapa samozřejmě nebude nikdy zcela dosažitelná, vždy bude obsahovat slepá místa. I jako částečně realizovaný budoucí objekt však vyvolává závažné etické otázky. I částečně hyperzmapovaný svět bude krmit prediktivní a autonomní systémy: než začneme ve světě působit, musíme ho nejprve vnímat. Třebaže nálepka „umělá inteligence" není vždy užitečná – technologie mytologizuje jako nový druh, jako autonomního morálního agenta mimo naši kontrolu –, algoritmy učení pročesávající hypermapu naše chápání světa nepochybně promění. Systémy počítačového vidění již dnes umějí předpovědět volební trendy z Google Street View, přičemž namísto zkoumání politického názoru používají poměr počtu pickupů a sedanů;[114] podobné systémy dnes docházejí i k jiným, škodlivějším závěrům.

[114] Timnit Gebru et al, „Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States", *Proceedings of the National Academy of Sciences*, listopad 2017.

Neo-fyziognomie

Stanfordští vědci Michal Kosinski a Yilun Wang v roce 2017 pronesli ohromující prohlášení: systém strojového učení údajně naučili na první pohled rozeznat homosexuální a heterosexuální osoby. Dosahuje prý 81% přesnosti u mužů a 74% u žen. Svá zjištění oba vědci spojili s kontroverzním názorem, že sexuální orientace má biologický základ v působení hormonů v prenatálním období.

Tváře homosexuálů jsou z hlediska genderu obvykle atypické. Průměrné obličejové rysy ukázaly, že gayové mají užší čelist a delší nos, zatímco lesby mají větší čelist. Kompozitní tváře naznačují, že homosexuální muži mají větší čela než heterosexuálové, zatímco lesby mají čelo menší než heterosexuální ženy.[115]



Kompozitní tváře heterosexuálů a homosexuálů, ze studie M. Kosinského a Y. Wanga. Přetištěno se svolením APA.

[115] From Michal Kosinski & Yilun Wang, „Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images", *Journal of Personality and Social Psychology*, únor 2018, sv. 114, č. 2.

V minulých staletích byla představa, že něčí obličej odhaluje jeho vlastnosti a charakter, známá jako fyziognomie; dnes se všeobecně odmítá jako bigotní pseudověda. Výzkum M. Kosinského a Y. Wanga – který rychle vešel ve známost jako „studie gaydaru" – vyvolal ostrou kritiku a kolegové i aktivisté ho odsuzovali jako pavědu a politováníhodné oživení fyziognomie. Někteří recenzenti studii napadli pro údajné metodologické nedostatky, například zjevné ignorování péče o vzhled a sebeprezentace, zatímco Greggor Mattson, sociolog z Oberlin College, odsoudil její „hermetickou odolnost vůči případným příspěvkům z oblasti sociologie, kulturní antropologie, feminismu a LGBT studií".

> [Kosinski a Wang] si myslí, že tvar nosu a lícních kostí jsou dané obličejové rysy; kdyby však potkali homosexuálního transvestitu, věděli by, že kontury lze měnit. Drží se neuvěřitelného předpokladu: že fotografie a profily na seznamkách jsou nezprostředkované, nemanipulované, přesné faksimile skutečného těla (stručně řečeno nejsou).[116]

Technologie pro rozpoznávání homosexuálů je vyloženě nebezpečná. Desítky zemí stále vězní lidi jen kvůli jejich sexuální orientaci; některé oběti bývají odsouzeny i k smrti. Extremisté na území ovládaném Islámským státem a Čečensko gaye zadržují, mučí a zabíjejí: není těžké si představit zvěrstva, jež by tato technologie způsobila v rukou homofobních radikálů.

Diagnostický a statistický manuál duševních poruch (DSM) uváděl homosexualitu jako poruchu až do roku 1973 a před Deweyho desetinným tříděním z roku 1996 se knihy o LGBTQ kultuře často řadily do oddílů „psychopatologie" nebo „sociální problémy".[117] Studie gaydaru představuje v sexuální klasifikaci krok zpět a ohrožuje celá desetiletí snah zařadit homosexuály mezi platné členy společnosti.

M. Kosinski je ve světě profilování všeobecně známý. Patří do týmu, který stál za experimentem s mikro-cílením společnosti Facebook zmíněným ve

[116] Greggor Mattson, „artificial intelligence discovers gayface. sigh.", *Scatterplot*, 10. září 2017, scatter.wordpress.com.

[117] Doreen Sullivan, „A brief history of homophobia in Dewey decimal classification", *Overland*, 23. července 2015, overland.org.au.

třetí kapitole, a má vazby na bezpečnostní firmu Faception, jež se na svém webu chlubí vlastní technologií rozpoznávání a ptá se: „Co kdyby bylo možné zjistit, zda je anonymní jedinec potenciální terorista, agresivní člověk nebo potenciální zločinec?" Ačkoli Kosinski tuto spojitost bagatelizuje, zjevně pro neo-fyziognomickou technologii vidí i jiné možnosti využití. Deník *The Economist* v propagačním článku o studii gaydaru informoval: „Podle Dr. Kosinského by se pomocí správných souborů dat mohly podobné systémy umělé inteligence naučit hledat jiné intimní vlastnosti, například IQ či politické názory."[118]

Autoři studii a své úmysly mocně hájili poukazem na to, že výzkum schválila etická komise Stanfordovy univerzity. To je pravda. Revizní komise často nebezpečí pramenící z projektů strojového učení podceňují, protože do nich nejsou přímo zapojeni žádní lidé; toto rozhodnutí etické komise je však velmi diskutabilní a ukazuje na nedostatek předvídavosti, pokud jde o případné zneužití dané technologie. Kosinski a Wang také zdůrazňovali, že jen manipulovali stávajícími technologiemi a nevytvářeli nic nového, a tvrdili, že výsledky zveřejnili proto, aby varovali před možnými riziky.

> Měli jsme pocit, že je nezbytně nutné, aby si zákonodárci a členové LGBTQ komunity uvědomili rizika, jimž čelí. Technologické firmy a vládní agentury si potenciál algoritmických nástrojů počítačového vidění uvědomují dobře. Domníváme se, že lidé si zaslouží tato rizika znát a mít příležitost podniknout preventivní opatření.[119]

„Když to neudělám já, udělá to někdo jiný"

Tím se dostáváme na území klasické etiky. Argument M. Kosinského je variantou tvrzení „Když to neudělám já, udělá to někdo jiný", běžné obhajoby těch, kteří pracují na kontroverzních projektech. Pokud se daná technologie

[118] „Advances in AI are used to spot signs of sexuality", *The Economist*, 9. září 2017, economist.com.

[119] Michal Kosinski a Yilun Wang, „Authors' note: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images", 28. září 2017.

vytvoří tak jako tak, bude lepší, když to udělají Kosinski a Wang než někdo méně eticky smýšlející?

Jde o poměrně solidní utilitární obhajobu. Podobné nástroje již nepochybně vytvořil nějaký pochybný režim nebo firma, takže možná bychom měli být Kosinskému za jeho údajný altruistický záměr upozornit veřejnost vděční: škoda na lidském štěstí může být o něco menší. Deontolog bude nicméně skeptický. Oslavovat vznik škodlivé technologie rozhodně správně není, ne? Mohli bychom sáhnout po generalizačním testu – Co kdyby to, co se chystám udělat, udělal každý? –, filozof David Lyons však vysvětluje, že otázku je třeba formulovat pečlivě. Svět by byl jistě horší, kdybychom například všichni začali pracovat pro zbrojařské společnosti, podle Lyonse by však spravedlivý generalizační test měl poukázat na to, že kdybychom takovou práci odmítli, přijal by ji někdo stejně schopný.

Randy Cohen, bývalý autor sloupku *New York Times* „The Ethicist" (Etik), obranu typu „Když to neudělám já..." nevybíravě kritizuje. „Slova ‚Když to neudělám já, udělá to někdo jiný' neospravedlňují hanebné chování. Někdo jiný udělá prakticky cokoli. ‚Někoho jiného' jsem potkal a je to malý podrazák."[120] Cohen má pravdu. Vlastní mravnost nemůžeme opírat o hypotetického jiného člověka: etická rozhodnutí musíme dělat sami za sebe. Tvrdit, že se něco stane s vámi, nebo bez vás, znamená smířit se se smutnou situací preventivní bezmocnosti. Kromě toho ospravedlněním „Když to neudělám já..." lze jistě omluvit jakýkoli hanebný čin.

> Pokud toto ospravedlnění přijmeme, těžko poznáme, které činy, ať již jakkoli podlé, by se nedaly obhajovat stejným způsobem. Roli nájemného vraha, regulátora přívodu plynu v Belsenu či hlavního mučitele z řad jihoafrické policie jistě někdo zastane, zdá se tedy, že to, jestli takovou práci přijmu, nebo odmítnu, nemá na celkový výsledek žádný vliv. – Jonathan Glover[121]

[120] Randy Cohen, „Doing the Outsourcing", *The New York Times Magazine*, 4. února 2011, nytimes.com.

[121] Jonathan Glover a M. J. Scott-Taggart, „It Makes no Difference Whether or Not I Do It", *Aristotelian Society Supplementary*, sv. 49 (1):171–209, 1975.

Větě „Když to neudělám já..." můžeme velkoryseji odporovat i poukazem na promarněné příležitosti. Jste-li dost duchapřítomní na to, abyste si kladli etické otázky, jistě můžete udělat něco pozitivnějšího než vymýšlet pochybné technologie, ne? Přenechte projekt dalšímu člověku v řadě; z morálního hlediska pravděpodobně odvede průměrnou práci, zatímco vy byste mohli udělat mnohem více dobra jinde. Signál, který principiálním odmítnutím vyšlete do společnosti, může mít navíc pozitivní vliv na vaše kolegy.

U jakéhokoli problematického projektu lze použít užitečný test: „Existuje situace, ve které je tato technologie prospěšná?" Žádné pozitivní využití technologie gaydaru nevidím. Tato technologie nabízí potenciálně smrtonosné vedlejší účinky a nulový přínos: jako čisté újmě by se jí mělo bránit už z principu. Nebezpečná technologie vytvořená jako varování je stále nebezpečnou technologií. Jak před dvěma tisíciletími řekl Bión z Borysthenu: „Chlapci házejí kameny na žáby pro zábavu, žáby však neumírají pro zábavu, ale doopravdy."

Může-li být nebezpečné třeba jen diskutovat o potenciálně škodlivé technologii, neplatí to i pro spekulativní design? Nehrozí riziko, že nebezpečné myšlenky normalizujeme a představíme právě těm, kteří je dokážou uvést do života? Hrozí. V květnu 2018 unikl z obvykle nekomunikativní dceřiné firmy „X" společnosti Alphabet morálně rozporuplný návrh, který v tisku vyvolal obavy.

> Co se stane, pokud se spekulativního designu zmocní korporátní společnost? Jestliže se praxe schovává za zdmi a dohodami o mlčenlivosti obřích firem v Silicon Valley, ztrácí postavení veřejné provokace a stává se něčím mnohem znepokojivějším. Nejvíc by nás mělo zneklidňovat, že se tak zaměstnancům společnosti umožní, aby promýšleli nemyslitelné – potýkat se s tím, jak vševědoucí a mocnou se tato právnická osoba může stát. – Felix Salmon[122]

Test potenciální prospěšnosti může být užitečným filtrem a dává nám další důvod k tomu, abychom se čistě spekulativním dystopiím vyhýbali.

[122] Felix Salmon, „The Creepy Rise of Real Companies Spawning Fictional Design", *Slate*, 30. května 2018, slate.com.

Tato epizoda by měla sloužit jako připomínka, že spekulativní design, provokatypy a designové fikce musí být řádně označeny jako fikce, aby se v případě úniku nezaměňovaly za skutečné plány, a musí je doprovázet patřičná etická diskuse. Designové fikce není dobré vypouštět do světa bez dozoru: cílem cvičení je výsledná debata, nikoli samotný artefakt.

Smrtící švy

Hypermapa se bude objevovat kousek po kousku a algoritmy fungující na základě její znalosti neožijí v dokonalé podobě. Automatizace bude nutně probíhat postupně. Rozdělení povinností mezi lidi a stroje může být bohužel velmi nebezpečné. Zvláště zrádné jsou okamžiky, kdy se předává štafeta: takzvané smrtící švy. Chris Noessel v knize *Designing Agentive Technology* (Navrhování agentních technologií) těmto štafetám věnuje celou kapitolu a říká jim „Achillova pata agentních systémů".[123] Podívejme se na smrtící švy v moderním kontextu: v tomto případě jimi jsou autonomní vozy.

K předání štafety může docházet v obou směrech. Předání kontroly počítači je obvykle menší problém: auto může oznámit, kdy je připraveno převzít řízení, v případě potřeby požádat řidiče o potvrzení přepnutí a potvrdit, že všechno proběhlo dobře. V úvahu je nutno vzít několik poruchových stavů, obvykle by však tento přechod měl probíhat pouze za stabilních, bezpečných podmínek. Naproti tomu předání kontroly člověku působí výrobcům autonomních vozů velké potíže. Nedávné údaje o experimentech s autonomními vozy v Kalifornii ukazují, že „vypnutí autopilota" (disengagement) se pohybuje od impozantní hodnoty 0,18 na tisíc mil (u vozů z projektu Waymo, považovaných za jasné tahouny v oboru) po alarmující hodnotu 755 (Mercedes-Benz).[124]

Hlavní nebezpečí při vypnutí autopilota představuje skutečnost, že autonomní vozy jsou pomalé a náchylné ke katastrofálním chybám. Akademici z University of Southampton zjistili, že řidičům trvalo 1,9 až 25,8 vteřiny, než po vypnutí autopilota znovu získali nad vozem plnou kontrolu; rozptýleným

[123] Chris Noessel, *Designing Agentive Technology* (Rosenfeld Media, 2017).

[124] State of California Department of Motor Vehicles, „Autonomous Vehicle Disengagement Reports 2017", dma.ca.gov.

řidičům to trvalo dokonce v průměru o 1,6 vteřiny déle.[125] Pozoruhodné přitom je, že první člověk, který na sedadle řidiče poloautonomního vozu zemřel, Joshua Brown, ignoroval opakované výzvy, aby držel ruce na volantu.

Letecký průmysl zná nebezpečí vypnutí autopilota až příliš dobře. Záhadná nehoda letadla Air France 447, jež roku 2009 spadlo do Atlantského oceánu, se nakonec vysvětlila jako řetězec pilotových chyb po nečekaném vypnutí autopilota. Autopilot letadla se při průletu silnou bouří sám vypnul, a letadlo tak přešlo do jiného režimu.

> Jakmile počítač ztratil údaje o rychlosti letu, autopilota odpojil a přešel z běžného režimu na „alternativní", který pilota omezuje mnohem méně. V alternativním režimu mohou piloti zabránit pádu letadla při ztrátě vztlaku pod křídly. – Jeff Wise[126]

Piloti si stav přístrojů vyložili špatně, novému režimu se nepřizpůsobili a přes slyšitelná varování trvali na tom, že zabránit ztrátě vztlaku pod křídly není možné. Letadlo spadlo kvůli lidské chybě, ovšem spirálu omylů odstartoval zpackaný přechod na lidské řízení.

Člověku, který přebírá řízení po autopilotovi, mohou chybět důležité informace, nebo může dokonce zapomenout, jak systém správně ovládat. Jak nás učil Marshall McLuhan, „každé rozšíření je také amputace".[127] Dovednosti řidiče obklopeného pohodlnou automatizací mohou časem upadat. Jestliže řidiči netráví za volantem dostatek času, nemohou se naučit lépe interpretovat situace, nezískají cit pro rytmus provozu a neprohloubí si znalosti pravidel silničního provozu.

Zvýšit bezpečnost smrtících švů není snadné. Podle jednoho silného argumentu by se řidičské zkoušky v poloautomatické éře měly zaměřovat spíše na vypínání autopilota, jenže trénovat uživatele nebude vždy možné. Některé projekty autonomních vozů zkoumají předávání řízení

[125] Alexander Eriksson & Neville A Stanton, „Take-over time in highly automated vehicles: non-critical transitions to and from manual control", *Human Factors* 59, 4, 2017.

[126] Jeff Wise, „What Really Happened Aboard Air France 447", *Popular Mechanics*, 6. prosince 2011, popularmechanics.com.

[127] Marshall McLuhan, *Understanding Media* (McGraw-Hill, 1964). (Česky: *Jak rozumět médiím: extenze člověka*. Přel. Miloš Calda. Praha: Odeon, 1991 – pozn. překl.)

dohledovému centru, kde nouzové situace na dálku zvládne vyškolený lidský operátor. Největší pozornost se však věnuje projektování bezpečného předání řízení. Noessel navrhuje pokud možno postupná preventivní varování: autonomní vůz by měl řidiče nejprve varovat před složitými podmínkami a pak mu oznámit, že se blíží limitům bezpečného řízení, a nechat ho, aby řízení manuálně převzal, nebo ho připravit na blížící se vypnutí autopilota. To však nebude možné v nouzových situacích, kdy musí systém, který se vzdává řízení, hrozbu vysvětlit neomylně, stručně a aniž by řidiče zaskočil, protože pak často zůstávají jako přimražení. Tento neobyčejný projektový úkol vyžaduje nalezení jemné, ale zásadní rovnováhy mezi příliš velkým a příliš malým množstvím informací. Konstruktéři se obvykle zaměřují na to, aby řidičovu pozornost poutalo vizuální upozornění, zvuk a hmatové vjemy, ovšem tak, aby nedošlo k jeho nebezpečnému přetížení.

Je lepší dost dobré?

Nedokonalá automatizace může být zrádná i bez smrtelného rizika. Stačí se zeptat policie z jižního Walesu, jež podle zákona o svobodném přístupu k informacím musela přiznat, že z 2470 možných podezřelých, na které upozornil jejich nový systém rozpoznávání obličejů, bylo falešně pozitivních 92 %.[128] Budete-li automatizovat bezpečnostní systémy, snažte se, aby fungovaly. Šetření BBC ukázalo, že bratr jednoho reportéra dokázal projít hlasovými identifikačními systémy a dostat se k jeho bankovnímu účtu.[129]

Kdy u částečně automatizovaného systému převažují výhody nad nedostatky? Stačí, aby technologie prostě předčila lidi, nebo by měly algoritmy splňovat vyšší standard? Jako východisko nám poslouží naše dva etické rámce. Nezapomeňte, že utilitaristu zajímají pouze důsledky, čistý dopad na lidské štěstí či utrpení. I nedokonalý automatizovaný systém může být spravedlivější, levnější a spolehlivější než lidé, a může tak přispět ke

[128] David Meyer, „Police Tested Facial Recognition at a Major Sporting Event. The Results Were Disastrous", *Fortune*, 7. května 2018, fortune.com.

[129] Dan Simmons, „BBC fools HSBC voice recognition security system", *BBC News*, 19. května 2018, bbc.co.uk.

zvýšení spravedlnosti a prosperity a zároveň omezit dřinu. Tyto výhody by se však měly porovnávat s potenciálními škodami i se současným stavem, jinými slovy s dopadem toho, že k zavedení dané technologie nedojde. Pomoci by mohla morální představivost s provokatypy a designovými fikcemi, stejně jako snaha brát v potaz širokou perspektivu zainteresovaných stran. Autonomní vozy například ovlivňují nejen řidiče, nýbrž také chodce, pojišťovny, urbanisty, a dokonce i životní prostředí, a to za možného předpokladu, že budou jezdit úsporněji.

Americký Národní úřad pro bezpečnost silničního provozu (NHTSA) definuje šest úrovní automatizace autonomního řízení.

	**Úroveň 0** Bez automatizace	Nulová autonomie, řidič vykonává všechny úkoly řízení.
	**Úroveň 1** Asistent řidiče	Vozidlo řídí řidič, ale konstrukce vozidla může zahrnovat některé funkce asistenta řidiče.
	**Úroveň 2** Částečná automatizace	Vozidlo má kombinované automatizované funkce, například akceleraci a ovládání volantu, ale řidič se musí věnovat řízení a neustále sledovat okolí.
	**Úroveň 3** Podmíněná automatizace	Řidič je zapotřebí, nemusí však sledovat okolí. Musí být vždy připraven po upozornění převzít kontrolu nad vozidlem.
	**Úroveň 4** Vysoká automatizace	Vozidlo za určitých podmínek zvládne všechny úkoly řízení. Řidič může mít možnost vozidlo ovládat.
	**Úroveň 5** Plná automatizace	Vozidlo zvládne všechny úkoly řízení za všech podmínek. Řidič může mít možnost vozidlo ovládat.

Úrovně automatizace podle NHTSA na základě původního rámce sdružení SAE (Society of Automotive Engineers).

Většina projektů autonomního řízení dnes běží na úrovni 2 nebo 3 (L2/3), liší se však kontext a podmínky. Auto typu L3 vytrénované na slunečných dálnicích jižní Kalifornie bude mít na moskevském prospektu plném sněhové břečky pravděpodobně co dělat: nějaký dopad bude mít počasí, ale hustota dopravy a neznámý místní styl jízdy mohou poloautonomní vozidlo přimět i k bázlivé pasivitě, která je sama o sobě riziková.

Podle utilitárního argumentu bychom měli vítat i poloautomatizaci úrovně L2/L3. Vzhledem k tomu, že lidská chyba přispívá k 90 % nehod,[130] každá technologie, která počet nehod omezí, přinese omezení škod. Společnost Tesla tvrdí, že po instalaci její aktualizace autopilota L2 se počet nehod snížil o 40 %, ačkoli toto číslo lze stěží ověřit. Nicméně již snížením počtu úmrtí na silnicích a počtu zranění (ohromné utilitární újmy) poloautonomní vozy téměř jistě zvýší kapitál štěstí.

U jiných automatizovaných systémů je utilitární argumentace pochybnější. Chytání zločinců pomocí rozpoznávání obličeje je sice dobré pro společnost, proti této kalkulaci však stojí ohromné množství falešně pozitivních výsledků. Falešně obviněny by mohly být tisíce lidí, což by ohrozilo poctivost policejní práce a naše pojetí samotné spravedlnosti. Tvrzení, že pravidla pro omezení přílišných pravomocí zaručují, že neprávem obviněná osoba bude obvinění rychle zproštěna, bude slabou útěchou pro Stevena Talleyho, jehož policie zadržela poté, co ho obličejový forenzní systém (třebaže ne na bázi softwaru) mylně identifikoval jako podezřelého z bankovní loupeže, a který kvůli policejní brutalitě při zatčení utrpěl trvalou ztrátu sluchu, čtyři zlomená žebra a poranění penisu a přišel o několik zubů.[131]

Deontologa více zajímají související morální povinnosti a mohl by namítnout, že každá nová technologie si zaslouží zvýšenou etickou pozornost. Vědomě pustit na trh poloautomatizované produkty, jež představují nové hrozby, ať už je jejich čistý dopad jakýkoli, pro něj bude těžší. Opravdu lze za správné považovat beta testování s lidskými životy? Neexistují lepší způsoby, jak dosáhnout stejných cílů, aniž bychom s lidmi jednali jako s prostředky k dosažení vlastního technologického pokroku?

[130] Bryant Walker Smith, „Human Error as a Cause of Vehicle Crashes", blog *The Center for Internet and Society*, 18. prosince 2013, cyberlaw.stanford.edu.

[131] Kirk Mitchell, „Man sues FBI and Denver police for $10 million claiming false arrest for 2 bank robberies and excessive force", *The Denver Post*, 15. září 2016, denverpost.com.

Vzhledem ke složitosti debaty není překvapením, že týmy vyvíjející autonomní vozidla zvolily mnoho různých přístupů. Účastníci projektu Waymo, šokovaní usínajícími testovacími řidiči ve vozech L2/3, jsou odhodláni skočit hned na úroveň L4, dříve než svou technologii představí veřejnosti. Společnost Tesla se naproti tomu rozhodla postupně aktualizovat software autopilota ve vozech, které již jezdí, a z jejich řidičů tak na jejich vlastní nebezpečí udělala stevardy veřejného testování autonomních vozidel.

Abychom to rozsekli, mohli bychom se utéct k Rawlsovu závoji nevědění. Možná je nejlepší poloautonomii řešit tak, jako bychom nevěděli, kde v systému skončíme. Byli bychom s projektem řízení na úrovni L2 stejně spokojení jako řidič, chodec nebo cyklista, regulační orgán či ekolog? A co s úrovní L4?

Díky roli licencování trh s autonomními vozidly efektivně funguje v souladu se zásadou předběžné opatrnosti. Společnosti potřebují k testování autonomních vozů na veřejných komunikacích souhlas – třebaže Uber testy prováděl bez povolení, jak má ve zvyku, a přinutil tak kalifornské úřady, aby je ukončily – a studie pak nemůže začít, dokud není regulační orgán spokojen. Technologie jako rozpoznání obličejů se tolik nekontrolují a mohou proklouznout mezerami v současném právu. Technologie počítačového vidění se běžně nejprve expedují a teprve pak přicházejí otázky. Tento přístup sice umožňuje rychlou inovaci, studie s gaydarem však naznačuje, že má i zhoubné vedlejší účinky. V současnosti sílí kampaň za regulaci, či dokonce zákaz rozpoznávání obličejů dříve, než bude pozdě.

> Představte si technologii, která je silně, jedinečně nebezpečná – něco v podstatě tak toxického, že si to zaslouží, abychom to naprosto odmítli, zakázali a stigmatizovali. Něco tak zhoubného, že regulace před účinky této technologie nemůže občany adekvátně chránit. Tato technologie již existuje. Jde o technologii rozpoznávání obličejů, která představuje tak velká nebezpečí, že ji musíme zcela odmítnout. – Evan Selinger[132]

[132] Evan Selinger, „Amazon Needs to Stop Providing Facial Recognition Tech for the Government", *Medium*, 21. června 2018, medium.com.

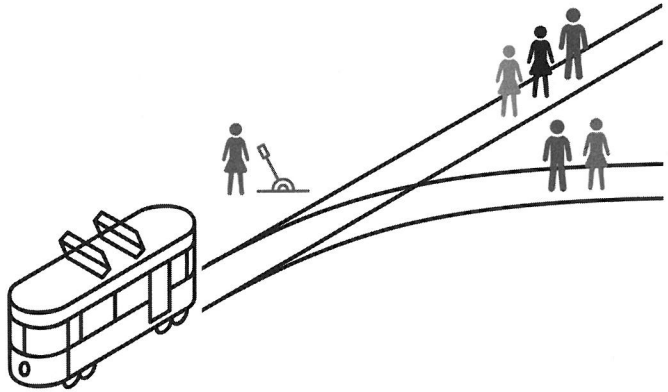
Regulační orgány budou pravděpodobně opatrné a technologickým firmám snad alespoň zakážou vyvolávat klamný dojem o možnostech automatizace. Autonomní systémy balí nové schopnosti do starého hávu: autonomní vůz vypadá jako běžné auto a kamera rozpoznávající obličeje vypadá jako běžná kamera, jde však o zařízení, jež se od těch předchozích zásadně liší. Veřejnost tedy musí s jistotou vědět, co systém smí a nesmí dělat. Rozhodnutí společnosti Tesla označovat svou technologii L2 za „autopilota" schopnosti systému nezodpovědně zveličuje; jde o triumf marketingu nad bezpečností. Povinností technologů je umožnit lidem vytvářet přesné mentální modely měnících se technologií. Řada zemí vyžaduje, aby dětské zbraně měly zářivou barvu, nebo alespoň oranžovou přední část hlavně, aby lidé poznali, že jde o hračku. Existuje dobrý etický důvod, proč by inteligentní a autonomní objekty měly být podobným způsobem odlišeny od svých inertních protějšků, a to buď designem, nebo etiketou.

Tramvajové dilema je zástupný problém

Nejvyšší úrovně automatizace autonomního řízení oživují slavný etický problém. V *tramvajovém dilematu*, které poprvé nadnesla Philippa Footová v roce 1967, si máme představit ujíždějící tramvaj, která se řítí na dělníky na kolejích.[133] Zastavit se dá jedině přepnutím výhybek, které ji sice přesměruje na novou trasu, ale zabije jiné dělníky. Přepnuli byste výhybku, která zabije dva dělníky, a zachrání tři? Většina lidí odpoví, že ano. Toto dilema má mnoho dalších variant. Abyste tramvaj zastavili a zachránili všechny, shodili byste z mostu tlouštíka? Většina lidí řekne ne, protože vidí morální rozdíl mezi „aktivním zabitím" a tím, když někoho „necháte zemřít". Dva dělníci, nebo jedno dítě? Zločinec, nebo učitel?

V budoucnosti autonomního řízení se z myšlenkového experimentu stává reálný problém: autonomní vozy si budou muset vybrat, zda zabrání jedné kolizi tím, že budou riskovat jinou. Tramvajové dilema jako takové se stalo ukázkovým příkladem technologické etiky, jež figuruje v desítkách novinových článků a objevuje se dokonce jako mem: Marx provádí „posun

[133] Philippa Foot, „The Problem of Abortion and the Doctrine of the Double Effect", *Oxford Review*, č. 5, 1967.



Tramvajové dilema: víte, co máte řešit.

na vícero kolejích", aby odstranil buržoazii; Camus se ke kolejím existenciálně přivazuje. Ve snaze porozumět postojům veřejnosti provedli na MIT crowdsourcingovou studii s názvem Morální stroj (Moral Machine). Zjistili, že lidé obecně aplikují utilitární logiku a rozhodují se pro takové dopady, jež by poškodily co nejméně lidí. Měla by se tedy autonomní vozidla řídit utilitárními principy?

Každé rozhodnutí autonomního řízení má tři kroky. Zaprvé musí vozidlo vnímat a identifikovat své prostředí, jak jsme uvedli v této kapitole. To však komplikují anomálie. Některé objekty jsou definované dobře – jízdní kolo vždy vypadá jako kolo, autobus jako autobus –, jenže některé jsou amorfní. Je tamto smetí, nebo pes? Dojde vašemu systému, že dívka v maškarním kostýmu zebry je stále člověk?

Pak musí systém předvídat, jak se objekty budou pohybovat. Chyba při identifikaci se přitom násobí: jakmile špatně klasifikujete objekt, nevydá se po očekávané trase. V březnu 2018 vůz společnosti Uber s úrovní autonomního řízení 2 v arizonském městě Tempe srazil a zabil chodkyni Elaine Hertzbergovou. Auto ji nejprve identifikovalo jako neznámý objekt, pak jako vozidlo a poté jako kolo, každá klasifikace přitom vedla k jiným odhadům trajektorie.[134] Vůz ani řidič – který měl sloužit jako pojistka a který několik

[134] Alexis C. Madrigal, „Uber's Self-Driving Car Didn't Malfunction, It Was Just Bad", *The Atlantic*, 24. května 2018, theatlantic.com.

minut před havárií streamoval pěveckou soutěž *The Voice* – až do srážky nebrzdili. V době, kdy jsem psal tento text, ještě probíhalo soudní jednání.

Posledním krokem v procesu je výběr a provedení akce. I když v tomto případě přichází teoreticky ke slovu tramvajové dilema, odpověď je obvykle prostá. Podle hlavního inženýra projektu Waymo Andrewa Chathama „je sice nutné odmyslet si určité intelektuální machinace, ale odpověď zní téměř vždy ‚dupněte na brzdu'".[135] S tímto přístupem je nepravděpodobný utilitární kalkul zbytečný. Pokud by měl autonomní vůz učinit utilitární rozhodnutí v reálném čase – snad si vzpomenete na utilitarismus konání –, musel by počítat s každou potenciální obětí, u každé odhadnout věk a povolání (účastníci studie Morální stroj raději chránili mladší a lépe situované osoby), zjistit pravděpodobnost a závažnost poranění a pak tyto škody zvážit u několika různých kroků: to všechno dříve, než se rozhodne jednat. Jako přijatelnější se jeví dát autonomnímu vozu spíše utilitární pokyny, soubor rozhodujících zásad, jejichž podstatou bude minimalizace škod, než chtít, aby vůz každou situaci promýšlel od začátku.

Utilitární přístup má ovšem ještě jednu chybičku: podle jistého článku v časopise *Science* by lidé raději jezdili v autech, jež upřednostňují jejich vlastní bezpečnost před bezpečností druhých. Účastníci studie nechtěli, aby vlády u autonomních vozidel prosazovaly utilitární logiku, a byli méně ochotni koupit si autonomní vůz, který by byl takto nastaven.[136] Přestože lze o tramvajovém dilematu psát pěkné úvahy, vzbuzuje toto dilema u veřejnosti také obavy, jež by zavedení autonomních vozů mohly bránit – zejména tu, že jako cestující můžete být obětováni pro větší dobro.

Zaměříme-li se na tramvajové dilema, odvádí nás to od závažnějších etických a sociálních otázek. Joanna Brysonová prozíravě poukazuje na to, že rozhodovat o tom, kdo při nehodě přijde k úrazu, můžeme již dnes: stačí si koupit SUV, a je dvakrát pravděpodobnější, že všechny sražené chodce zabijete.[137] Podle etika Johna Danahera může toto dilema také zastírat

[135] Alex Hern, „Self-driving cars don't care about your moral dilemmas", *The Guardian*, 22. srpna 2016, theguardian.com.

[136] Jean-François Bonnefon et al, „The social dilemma of autonomous vehicles", *Science*, sv. 35, 2016.

[137] Eric D. Lawrence et al, „Death on foot: America's love of SUVs is killing pedestrians", *Detroit Free Press*, 1. července 2018, freep.com.

riziko podjatosti. Má-li vaše autonomní auto potíže s rozpoznáváním černochů, srazí více černochů. Danaher přichází s překvapivým tvrzením, že dopady strukturální diskriminace by mohla omezit mírná randomizace.[138]

Autonomní řízení bude mít významné dopady na města, nikdo se však bohužel neshodne na tom jaké. Mnohé urbanisty nejistota ochromuje: blíží se nová éra dopravy, nebo se její potenciál přeceňuje? Obchodní spekulanti s portfoliovými AV společnostmi prosí guvernéry, aby pozastavili výdaje na veřejnou dopravu a počkali na blížící se rozmach autonomních vozů, ovšem kdy k němu dojde, je popravdě záhada. Jistě víme jen to, že pokud autonomní vozy dostojí svému potenciálu, dojde k prudkému poklesu dárcovství orgánů: 20 % všech darovaných orgánů v USA pochází od obětí nehod. Cesta před námi je plná nezamýšlených důsledků.

Soužití a společníci

Vzhledem prvním náznakům hnutí použitelnosti se sami designéři často zabývají efektivitou uživatele. S nástupem automatizace však nebudeme technologie ani tak používat jako s nimi koexistovat. I dnes lidé překvapivě prahnou po spojení s tzv. umělými agenty:

> Vždy jsem předpokládala, že chceme zachovat určitý odstup mezi námi a umělou inteligencí, ale zjistila jsem, že opak je pravdou. Lidé jsou ochotni navazovat s umělými agenty vztahy za předpokladu, že jsou sofistikovaně konstruované a schopné komplexní personalizace. Zdá se, že si my lidé chceme uchovat iluzi, že umělé inteligenci na nás záleží. – Liesl Yearsleyová[139]

Selže-li koncept uživatele, selže i na něj orientovaný design. Nově vznikající technologie nejsou pouhé produkty či nástroje: často jde o fyzické subjekty, které jednají částečně autonomně – jinými slovy o roboty. Designéři

[138] John Danaher, „The Ethics of Crash Optimisation Algorithms", *Philosophical Disquisitions*, 28. dubna 2017, philosophicaldisquisitions.blogspot.co.uk.

[139] Liesl Yearsley, „We Need to Talk About the Power of AI to Manipulate Humans", *MIT Technology Review*, 5. června 2017, technologyreview.com.

musí plánovat nejen, jak vypadají a jak se chovají, ale i to, jaké budou mít ve společnosti postavení. Tropy ze science-fiction mohou být v tomto případě limitující. Jak tvrdí Matt Jones ze společnosti Google, je až příliš snadné vehnat roboty do stávajících mocenských vztahů a udělat z nich infantilizované klony lidí či jakýsi patolízalský poddruh prahnoucí po uspokojování lidských vrtochů. Technologové jiná než lidská stvoření využívají ve jménu pokroku běžně: vzpomeňte na nebohou Lajku umírající na oběžné dráze či na Thomase Edisona a Harolda Browna Pitneyho, kteří elektřinou zabíjeli toulavá zvířata, aby ukázali nebezpečí střídavého proudu. Nové formy vztahu mezi člověkem a robotem by nám mohly umožnit se těchto nešťastných hierarchií a s nimi souvisejícího zneužívání zbavit. Nemohli by roboti člověka doprovázet spíše v symbióze než jako nevolníci?

Donna Harawayová se v knize *The Companion Species Manifesto* (Manifest mezidruhových společníků) rázně a provokativně dívá na skutečné soužití dvou druhů, lidí a psů, kteří spolu mají „závazný, konstitutivní, historický, ‚proměnlivý' vztah".[140] Jejím poselstvím je objev, že ke skutečnému kamarádství patří pocit, že se v něm daří oběma, vztah vzájemné odpovědnosti a respektu. Podstatou je prý poznání, že jiný druh není jako my: tento rozdíl bychom měli uznat a přijmout a měli bychom pochopit, že i přesto jsou jiné druhy důležité bytosti. Harawayová přesvědčivě tvrdí, že právě naladění na potřeby Jiného nám umožňuje pochopit, jak lze přispět ke vzniku světa snesitelnějšího pro všechny a pro všechno. Není to vlastně podstata etiky?

> Domnívám se, že veškeré etické vztahování, v rámci jednoho druhu či mezi druhy, je spleteno ze silného hedvábí neustálé ostražitosti vůči jinakosti-souvztažnosti. Nejsme jedno a bytí závisí na tom, jak spolu vycházíme. – Donna Harawayová[141]

[140] Donna Haraway, *The Companion Species Manifesto* (University of Chicago Press, 2. vyd., 2003).

[141] Ibid.

Osvětí

Jiné druhy zažívají svět, který by lidé nepoznali. Můžeme si sice myslet, že spektrum zahrnuje barvy od červené po fialovou, ovšem včely, omámené ultrafialovým zářením, je znají jinak. Losos zase dokáže vnímat magnetické pole Země a využívat ho při plavbě ohromnými oceány. Biolog Jakob von Uexküll v roce 1909 navrhl pro označení celého světa tak, jak ho vnímá živý tvor, pojem *osvětí* (Umwelt). Osvětí bytosti závisí na jejích smyslech. Svět psa tvoří pachy a zvuky; noční osvětí netopýra se točí kolem ultrazvuku. Každý druh žije ve vlastní informační podmnožině téhož světa, a pokud je schopen si něco uvědomovat, svět *je* to, co může poznat. Představa, že svět zahrnuje množství nepostřehnutelných informací, by nám měla být povědomá z diskuse o neviditelném světě dat, hypermapě a extrakčních schopnostech například počítačového vidění.

Díky osvětí můžeme o soužití přemýšlet jinak. O osvětí společníků se již lidé starají: jsme shovívaví ke psům, když strčí čenich do strouhy; život primátů v zajetí obohacujeme taktilními hračkami. Nebyla by pro každou skupinu dobrým principem designu pro soužití snaha respektovat a zlepšovat svět, v němž žije ten druhý?

Roboti mohou lidské osvětí obohatit zhmotněním neviditelného, převodem věcí, které mohou vnímat jen oni, do formátu čitelného pro lidi. Z toho vyplývá, že oni sami by měli být čitelní. Chceme-li žít po jejich boku, měli bychom dokázat porozumět jejich chování a mít jasno v tom, kdo odpovídá za jejich činy.

Soužití by samozřejmě mělo být oboustranné. Osvětí robotů určují lidé, alespoň prozatím. (Filozofové by mohli remcat, že aby mohl robot mít osvětí, potřebuje něco jako mysl, či přinejmenším smysly, ale chápete, jak to myslím). Lidé rozhodují o tom, které snímače se robotům nainstalují a jakými daty se nakrmí algoritmus. I my můžeme zhmotnit informace, jež jsou robotům jinak neviditelné, a to vytvořením světa, který bude čitelný pro ně. Měli bychom jim poskytnout bohatá, nezkreslená tréninková data, aby mohli vytvářet přesné modely a dokázali je rychle opravit, pokud budou vycházet z mylných předpokladů nebo chyb. Jestliže se rozhodneme vyvíjet roboty coby společníky, nejen coby pouhé nástroje, měli bychom toto rozhodnutí vnímat jako morální; možná že snaha rozšířit osvětí je téměř aktem milosrdenství vůči strojům. Pokud tomu tak je,

již zmíněné etické problémy s daty jsou zásadní nejen pro rozvoj člověka, ale i rozvoj stroje.

Musíme si však uvědomit, že teď jsme vyšli z dalekosáhlého a poněkud sentimentálního předpokladu: totiž že rozkvět umělé inteligence bychom měli vítat. Šestá kapitola nás od tohoto názoru spíše odradí.

Společenská smlouva

Je známo, že filozof Thomas Hobbes v 17. století viděl lidskou povahu pesimisticky. Soudil, že bez politické autority by lidé existovali v čistém „přírodním stavu"; těšili by se sice neomezené volnosti, ale zažívali by neustálý konflikt.

> V takovém stavu není místo pro průmysl; protože jeho plody jsou nejisté [...] žádné umění, žádná literatura, žádná společnost a, co je nejhorší, jen neustálý strach a nebezpečí násilné smrti a osamělý, chudý, ohavný, zvířecí a krátký život. – Thomas Hobbes[142]

Abychom tuto dystopii odvrátili, uzavíráme mezi sebou nevyslovený závazek, kterému dnes říkáme *společenská smlouva*. Ochotně se podřizujeme autoritám v podobě vlád, monarchií či soudů a některé svobody vyměňujeme za zabezpečení a sociální stabilitu. Podle Hobbese je tato smlouva základem politické legitimity, přičemž alternativou je chaos a nemorálnost. Anarchisté by však zřejmě s tím, že je autorita nezbytná, nesouhlasili, a altruismus je navzdory údajné společenské smlouvě bohatě dokumentován i v živočišné říši.

Myšlenka společenské smlouvy je nicméně politicky i eticky užitečná. Jeden z možných problémů této smlouvy řeší Rawlsův závoj nevědění: všichni bychom chtěli, aby byla napsána v náš prospěch. Závoj nevědění používáme, aby inteligentní lidé neobhajovali pravidla, jež straní inteligenci, a bohatí nehledali taková, jež zvýhodňují bohaté, a také se rozhodujeme pro zákony, jako bychom neznali okolnosti ani své osobní charakteristiky. Je jen

[142] Thomas Hobbes, *Leviathan*. (Česky: *Leviathan aneb Látka, forma a moc státu církevního a politického*. Přel. Karel Berka. Praha: OIKOYMENH, 2009 – pozn. překl.)

spravedlivé, že s pravidly hry souhlasíme ještě předtím, než se všem hráčům rozdají karty.

Zatímco každodenní život prostupují tržní síly, společenská smlouva spočívá čím dál častěji v rukou byznysu. Důležitým nepsaným ujednáním je například placení daní. Naše společné prostředí a instituce se musí nějak financovat, technologičtí giganti však navzdory snižování daní drží ohromné peněžní rezervy v zahraničí, mimo dosah úřadů. Toto jasné vykrucování se z povinností společenské smlouvy zůstává bohužel bez trestu: díky monopolní síle je snazší se závazkům vyhýbat.

Společenská smlouva spoléhá na reciprocitu: smlouva funguje, pouze když všichni souhlasíme s jejími podmínkami. Navrhovat soužití tedy částečně znamená navrhovat je pro důvěru a vzájemný prospěch. Měli bychom vytvářet jenom takové technologie, jež umožní rozvoj lidstva, a na oplátku bychom měli umožnit rozkvět technologií.

Vysvětlitelné algoritmy

Chceme-li důvěřovat, musíme rozumět. Proto je nutné, aby některé systémy strojového učení uměly objasnit svá rozhodnutí; tento princip je známý jako vysvětlitelná umělá inteligence (XAI).

Vysvětlování, jak rozhodujeme o druhých, je eticky rozumnou i důležitou podmínkou demokracie. Policie musí při zatýkání obvykle uvádět pravděpodobné důvody zatčení a soudy musí vysvětlovat veškeré výsledné rozsudky. Je také dokázáno, že vysvětlitelné technologie zlepšují porozumění na straně uživatele. Studie ladění chyb s vysvětlením postupu, při němž algoritmus uživatelům sdělil, jak dochází k předpovědím, ukázala, že její účastníci opravovali chyby systému dvakrát efektivněji.[143]

V cestě však bohužel stojí matematika. Některé metody strojového učení, systémy hlubokého učení (tedy nové obdoby neuronové sítě, ovšem s mnohem větším grafickým procesorem) a genetické algoritmy jsou už z podstaty neprůhledné. Prosté zveřejnění kódu nefunguje: veřejnost kódu nerozumí a nejde o spustitelné programy v klasickém smyslu. Hluboké

[143] Todd Kulesza et al, „Principles of Explanatory Debugging to Personalize Interactive Machine Learning", *Proceedings of IUI*, 2015.