UNIVERZITA PALACKÉHO V OLOMOUCI

FILOZOFICKÁ FAKULTA

INTERDISCIPLINÁRNÍ SPOLUPRÁCE V ZÁJMU VÝVOJE ETICKÉ A SPOLEČENSKY PROSPĚŠNÉ UMĚLÉ INTELIGENCE

Magisterská diplomová práce

Olomouc 2024 Ing. Lucie Stewart

UNIVERZITA PALACKÉHO V OLOMOUCI

FILOZOFICKÁ FAKULTA

KATEDRA SOCIOLOGIE, ANDRAGOGIKY A KULTURNÍ ANTROPOLOGIE

INTERDISCIPLINÁRNÍ SPOLUPRÁCE V ZÁJMU VÝVOJE ETICKÉ A SPOLEČENSKY PROSPĚŠNÉ UMĚLÉ INTELIGENCE

Magisterská diplomová práce

Studijní program: Management vědy

Autor: Ing. Lucie Stewart

Vedoucí práce: Mgr. Tomáš Dvořák, Ph.D.

Olomouc 2024

Prohlašuji, že jsem magisterskou diplomovou práci na téma *„Interdisciplinární spolupráce v zájmu vývoje etické a společensky prospěšné umělé inteligence“* vypracovala samostatně a uvedla v ní veškerou literaturu a ostatní zdroje, které jsem použila.

V Olomouci dne. 30. 3. 2024 Podpis ……

Ráda bych tímto vyjádřila poděkování vedoucímu diplomové práce Mgr. Tomáši Dvořákovi, Ph.D. za jeho cenné rady a postřehy při zpracování tématu práce. Velmi děkuji vedoucímu Katedry počítačů na FEL ČVUT doc. Ing. Jiřímu Vokřínkovi, Ph.D., vedoucímu Centra umělé inteligence prof. Dr. Michalu Pěchoučkovi, MSc., a tajemníku katedry Mgr. Jaroslavu Šípovi za konzultace a podporu během studia. Děkuji kolegům a kolegyním z Kateder počítačů a kybernetiky FEL ČVUT, kteří mi v průběhu studia byli vždy ochotní dát zpětnou vazbu.

V neposlední řadě děkuji své rodině za její podporu při studiu.

Obsah

Anotace

Jméno a příjmení:	*Ing. Lucie Stewart*
Katedra:	Katedra sociologie, andragogiky a kulturní antropologie
Studijní program:	*Management vědy*
Studijní program obhajoby práce:	*Management vědy*
Vedoucí práce:	*Mgr. Tomáš Dvořák, Ph.D.*
Rok obhajoby:	*2024*

Název práce:	*Interdisciplinární spolupráce v zájmu vývoje etické a společensky prospěšné umělé inteligence*
Anotace práce:	Cílem práce je zodpovědět na výzkumnou otázku: Jaké kontrolní mechanismy ovlivňují výzkum a výstupy v projektech využívajících nebo vyvíjejících umělou inteligenci, aby byly v souladu s etickými principy? Potřeba nalézt odpověď na tuto otázku vyplývá z praxe autorky, která je manažerkou výzkumných projeků na Katedře počítačů ČVUT FEL. Pokrok ve výzkumu umělé inteligence (UI) , vývoji a aplikaci UI v rámci spolupráce s dalšími vědeckými obory, výzkumnými oblastmi a tématy, je velmi rychlý. Přináší nové příležitosti, ale i hrozby, které je ve výzkumu potřeba reflektovat, a to nejen po technické stránce a zkoumáním společenských dopadů těchto technologií, ale i z hlediska managementu vědeckých projektů, příležitostí zapojení se do mezinárodních výzkumných konsorcií a partnerství, a výzev, které to klade na nejvyšší management výzkumných organizací, interní procesy a schopnost flexibilně reagovat na požadavky v oblasti vzdělávání, kontroly, zajištění, posílení a podpory vědecké integrity a management rizik a aktivní zapojení se do mezinárodních a národních struktur a organizací. Práce vychází z teorie v oblasti etiky UI a jejího postupného přenosu do reálného vědeckého prostředí – v rámci empirického výzkumu (rozhovory) a později aplikací v praxi.

	O etické otázky vědci zájem mají a nečekaným pozitivním výstupem tohoto výzkumu bylo založení neformální skupiny Responsible AI, která diskutuje témata spojená s etikou UI v rámci pravidelných schůzek. Praktickým výstupem je potom lepší schopnost autorky zapojit se do přípravné fáze vědeckých projektů (tzv. pre-award) právě v oblasti evaluace etických otázek spojených s výzkumem a vývojem UI, impaktu a prevence rizik.
Klíčová slova:	Etika; umělá inteligence; interdisciplinární spolupráce; výzkum, vývoj, inovace; management vědy; VaVaI; odpovědný výzkum a inovace, RRI
Title of Thesis:	Interdisciplinary cooperation with the aim of developing ethical and socially responsible artificial intelligence
Annotation:	The main goal of this diploma thesis is to answer the research question: Which control mechanisms influence research and outcomes of the research projects in the field of Artificial Intelligence (AI) in order to make sure the ethical principles are fully respected. The author works as a project manager and administrator (RMA) at the Department of Computer Science, at CTU FEE in Prague, thus the interest and choice of the research question. Navigating the rapid progress in the field of AI research, its development and applications in cooperation with other research disciplines is challenging. There are many opportunities but also threaths that need to be reflected in the responsible approach to research in order to find technical solutions in line with the ethical principles, and to learn and understand social dimension and impact of the AI. There are also challenges from the point of view of the research management and new opportunities to join international research consortia and partnerships, as well as the challenges for the top management of research institutions, internal processes and their ability to react accordingly to the requirements in the field of education, management and control, support of responsible research and innovation and risk

	management, and active participation in the international and national research support structures and organisations. The author's theoretical study in the field of AI ethics led to a gradual transfer of this topic into the research environment at the Department in the form of interviews and discussions about AI ethics with the researchers and, as a practical outcome, an informal Responsible AI group was set up. The researchers are interested in the ethical dimension of their work and some of them regularly meet to discuss many related topics and current issues within the group. Another practical outcome is an improvement of author's knowledge and experience with the AI ethics field in order to provide better support to the scientists during the pre-award stage of the grant proposal application process to eliminate potential risks and increase positive impact of their research.
Keywords:	Ethics; artificial intelligence; interdisciplinary cooperation; Research, Development and Innovation; science management, Responsible Research and Innovation (RRI)
Názvy příloh vázaných v práci:	
Počet literatury a zdrojů:	144
Rozsah práce:	*128* s. (*228 345* znaků s mezerami)

Úvod

Cílem práce je zodpovědět výzkumnou otázku: Jaké kontrolní mechanismy ovlivňují výzkum a výstupy v projektech využívajících nebo vyvíjejících umělou inteligenci, aby byly v souladu s etickými principy?

Potřeba nalézt odpověď na tuto otázku vyplývá z praxe autorky, která je manažerkou výzkumných projeků na Katedře počítačů ČVUT FEL. Pokrok ve výzkumu umělé inteligence (UI), vývoji a aplikaci UI v rámci spolupráce s dalšími vědeckými obory, výzkumnými oblastmi a tématy, je velmi rychlý.

Přináší nové příležitosti, ale i hrozby, které je ve výzkumu potřeba reflektovat, a to nejen po technické stránce a zkoumání společenských dopadů těchto technologií, ale i z hlediska managementu vědeckých projektů, příležitostí zapojení se do mezinárodních výzkumných konsorcií a partnerství, a výzev, které to klade na nejvyšší management výzkumných organizací, interní procesy a schopnost flexibilně reagovat na požadavky v oblasti vzdělávání, kontroly, zajištění, posílení a podpory vědecké integrity a management rizik a aktivní zapojení se do mezinárodních a národních struktur a organizací.
Vycházela jsem z následujících hypotéz:

Hypotéza 1: Propojení oborů UI a etiky zatím zcela chybí, protože se nepropisuje do vědecké praxe ani managementu projektů, kde je UI předmětem výzkumu nebo je využívána v kontextu jiných oborů a dopadů v celé řadě oblastí a profesí.

Hypotéza 2: Vývoj UI, která bude společensky prospěšná a v souladu s etickými principy je možné zajistit interdisciplinání spoluprací s filosofy a etiky v rámci spolupráce na společných projektech.

Hypotéza 3: Lze nalézt souvislost mezi kontrolními mechanismy a zlepšením kvality vyvíjených systémů UI v oblasti etiky a pozitivního impaktu na společnost?

Tato magisterská práce neřeší detailně konkrétní technické aspekty výzkumu a vývoje etické a společensky prospěšné umělé inteligence (tzv. technickou governance), ani neprovádí hloubkovou legislativní analýzu např. Nařízení EU v oblasti umělé inteligence (tzv. AI Act). To jí neumožňuje rozsah ani specializace autorky. Nejedná se ani o čistě filosofickou práci, i když právě z filosofie koncepce principů etické UI vychází. Všechny tyto výzkumné oblasti jsou však v DP do určité míry zastoupeny, nebo byla jejich znalost předpokladem k pochopení souvislostí. Praktickým výstupem je potom lepší schopnost autorky zapojit se do přípravné fáze vědeckých projektů (tzv. pre-award) právě v oblasti evaluace etických otázek spojených s výzkumem a vývojem UI, impaktu a prevence rizik.

1 Výzkum a vývoj umělé inteligence (UI)

Věda, nové technologie a design měly vždy přímý vliv na chod a rozvoj společnosti a měnily některé zvyky, tradice, procesy, kvalitu života, způsob seberealizace, trávení volného času a získávání prostředků k přežití.

Umělá inteligence (UI) je na jednu stranu nehmotná. Existuje ve formě algoritmů, pomyslných nul a jedniček. Na druhou stranu může být fyzicky přítomná, díky ztělesnění[1] v podobě robotů, průmyslových či asistenčních, nebo v oblasti zábavy. Má obrovský aplikační potenciál a už dnes je využívána prakticky ve všech sférách lidského života, ve výrobě, medicíně, zábavním průmyslu, vzdělávání a komunikaci.

Pro ukázku, jak komplexní systémy UI jsou a jak jsou provázané, uvádím příklad nasazení UI v průběhu pandemie Covid-19 v letech 2020-2022, obrázek

[1] Angl. termín z robotiky je embodyment.

(č. 1) z vědeckého žurnálu Nature Medicine (Budd et al, 2020). Selhání jednoho z těchto článků, vede k chybovosti zbytku systému.

Obrázek 1: Využití UI v aplikacích, výzkumu, testování, prevenci, diagnostice, komunikaci a dalších opatřeních při pandemii Covid-19

Tato vzájemná závislost může být jedním z možných rizik. Na druhou stranu, bez mobilizace a multidisciplinární spolupráce vědců technických, konkrétně UI komunity, virologů, epidemiologů, lékařů, psychologů a dalších oborů, by nebylo možné negativní dopady pandemie minimalizovat. Už jen nasazení UI v oblasti komunikace, sdílení a vyhodnocování dat, umožnilo vědeckým týmům pracovat na dálku, ať už jejich členové byli v tu dobu kdekoliv. Nasazeny byly i nejrůznější pipetovací roboty, drony a další technika.

I přes evidentní ztráty a tragédie, ke kterým došlo, byla tato vědecká „mobilizace" v tak krátkém časovém horizontu nevídaná a lze ji počítat mezi úspěchy vědecké komunity. V momentě krize byly všechny tyto obory schopné spolupracovat na společném cíli. Vyšly však i kritické studie, např. článek v MIT Technology Review v roce 2022, který vycházel ze studie Institutu Alana Turinga. Institut zkoumal 415 publikovaných nástrojů využívajících UI, natrénovaných na pacientských datech (konkrétně skenů plic z počítačové

tomografie a RTG snímky hrudníku u pacienů s příznaky onemocnění Covid-19) prostřednicvím využití metody strojového učení.[2]

Výsledkem studie je, že ani jeden z těchto nástrojů nebyl v dané situaci dostatečně vhodný/připravený pro využití v klinické praxi (Heaven, 2022). Hlavním důvodem byla nízká kvalita vstupních dat, na kterých byla UI trénována. Některé datasety se při trénování objevovaly několikrát (tzv. „Frankensteinovy datasety", tedy data slepená z více zdrojů), takže v konečném důsledku způsobovaly falešnou pozitivitu, bias, těchto nástrojů. Naléhavost nalézt rychlé řešení uprostřed pandemie způsobila i využívání příbuzných datasetů, tedy pediatrických pacientů, místo dospělých. Jednalo se v tomto případě o diagnostické nástroje ke stanovení diagnózy Covid-19, které kvůli zkreslení vstupních dat nefungovaly přesně se všemi možnými následky (Roberts et al, 2021).[3]

Dalším projektem, který podporuje Evropská unie, a dokazuje, jak komplexní je nasazení generativní UI napříč klíčovými průmyslovými odvětvími, je iniciativa GenAI4EU.[4] Generativní UI je technologie UI, „která může vytvářet různé typy obsahu, včetně textu, obrázků, zvuku či syntetických dat." Učením dokáže z trénovacích dat vytvářet data nová (Barták, 2024). Cílem

---

[2] **Strojové učení**, anglicky machine learning, je způsob trénování umělé inteligence založený na velkém množství vstupních dat, ze kterých si UI sama vyvozuje vzájemné vztahy, vazby a kontext. Využívají se algoritmy, které upravují vztahy, postupy a priority/cíle. Často vzniká fenomén tzv. černé skříňky – black boxu – kdy není možné zpětně vystopovat, jak k danému výsledku UI dospěla, z čeho čerpala a jak vstupní data vyhodnotila, aby jí tento výsledek vyšel. Dalším negativním výstupem strojového učení jsou různé formy zkreslení (bias) a diskriminace (Holzinger et al., 2018, s. 1-8). S rozklíčováním algoritmického biasu pomáhá např. metoda algoritmického auditu, kterým se zabývá např. Meredith Broussard (knihy Artificial Unintelligence a More Than a Glitch). U určitých oblastí diskriminace je potřeba nejdříve změnit společnost a její přístup k těmto otázkám spíše než čistě technické řešení pomocí více a kvalitnějších dat (Corbyn, 2023).

[3] Světová zdravotnická organizace (WHO) na tento problém zareagovala tak, že v případě krizí zvažuje zavedení krizového sdílení dat. To by umožnilo vědcům data jednodušeji globálně sdílet. Nazývají to „datovou připraveností." WHO (Heaven, 2022, August 2; Roberts, et al. 2021).

[4] Vizi a cíle tohoto projektu představil v rámci kick-off konference evropského projektu (OP JAK) ROBOPROX pořádaném ČVUT v Praze 14. 3: 2024, člen Zastoupení Evopské komise v České republice, Josef Schwarz. Stránky projektu ROBOPROX zde: https://roboprox.eu/

je stimulovat co nejširší aplikaci UI srkz 14 stěžejních průmyslových odvětví EU, jako například: letectví, mobilita, udržitelnost životního prostředí, zpracovatelský průmysl, robotika, zdravotnictví, biotechnologie, baterie a materiály. V projektu se počítá s inter/multidisciplinární spoluprací napříč těmito odvětvími, ale i mezi různými aktéry v rámci celé ekonomiky, od univerzit, přes startupy, průmyslové podniky a korporace až po veřejný sektor.

Způsob, jakým UI mění současnou společnost ve všech oborech a oblastech, však je často skrytý. Změny ve společnosti a vliv, který různé aplikace využívající UI mají na každého z nás, nejsou vždy hned rozpoznatelné a předvídatelné. Některé jsou a budou patrné až v dlouhodobém časovém horizontu.

Tento negativní dopad může být výsledkem biasu popsaným výše, tedy nereprezentativními datovými sadami, nedomyšlených souvislostí, kompromisy v době krize, technického selhání, naivity či nedůslednosti konečného uživatele, chyby, v některých případech však budou důsledkem i záměru jeho tvůrců. Většina známých případů zneužití umělé inteligence např. v politickém kontextu, jako případ firmy Cambridge Analytica, která využila neetickým způsobem osobní data z profilů uživatelů sociálních sítí k zacílení politické kampaně ve Spojeném království (Brexit) a v prezidentské kampani USA, byla unikátní a před existencí této technologie nebylo zřejmé, že k podobné situaci dojde (Amnesty International, 24. července, 2019). Jak tedy podobné situace ošetřit, jak jim předcházet a jak správně posoudit odpovědnost jednotlivých aktérů v případě zneužití?

Nesou potom odpovědnost i vědci a vývojáři, investoři, samotná firma nebo politik, který si kampaň objednal? Komplexní je např. debata kolem zodpovědnosti z provozu autonomního vozidla v případě havárie a újmy na životě či majetku. Kdo je zodpovědný? Má mít řidič možnost a zodpovědnost vůz v podobné situaci ovládat? Bude mít dostatek zkušeností s řízením, když

většinu času řídit prakticky nebude?[5] Takové možnosti či přímo podmínce zásahu člověka do činnosti systému UI se říká human-in-the-loop (HITL). Operátor (např. řidič, vojenský velitel zásahu atd) zůstává v tzv. kontrolní smyčce. Může tedy kdykoliv během operace prováděné autonomním systémem UI zasáhnout a činnost systému zastavit.

Výše uvedené jsou příklady aplikovaného výzkumu, kdy se univerzity zapojují do různých konsorcií a partnerství financovaných místními vládami, agenturami jako TA ČR, nebo přímo Evropskou komisí. Aplikovaný výzkum však probíhá i přímo na zakázku ve spolupráci s výrobními podniky a firmami. V oblasti základního výzkumu však jeho dopady nejsou dopředu zřejmé, jak vyplývá z rozhovorů, které jsem vedla s některými vědci a studenty doktorského studia na ČVUT FEL. Zde tedy žádná přímá odpovědnost výzkumníků za výsledky a využití jejich práce není? Jak tuto oblast ošetřit, jaká pravidla zvolit, pokud jsou vůbec nutná? (viz Vzdělávání a odpovědnost, podkapitola 3.3)

1.1 Vymezení základních pojmů

> *„Digitální revoluce […] se děje právě nyní. Jsme poslední generace, která zažila čistě analogový, před-digitální život mimo internet* (Floridi, 2023, s. xi).“

Tato podkapitola vysvětluje základní termíny, se kterými pracuje následující text: inteligence, umělá inteligence, základní a aplikovaný výzkum a interdisciplinarita. Nejedná se o pouhé definice, ale i o zamyšlení nad obsahem a historií daného termínu – zde konkrétně inteligence a umělé inteligence.

[5] Vlastní rozhovor s J. Hvoreckým z CETE-P, Filosofického ústavu AV ČR, 9. 1. 2023.

1.1.1 Inteligence

„Pojem ‚inteligence' u živých tvorů nebyl přesně vymezen. Existují sice metody ‚měření' inteligence (např. test IQ), ale každá z nich má svá omezení a nikdo nemůže tvrdit, že některá z nich jsou zcela objektivní" (Mařík, 1993, 16). Dalšími filosoficko-psychologickými pojmy, které se k UI vztahují a je velmi obtížné je přesně definovat, jsou: vědomí, bytí nebo svobodná vůle (Ibid.).

Cave (2020, s. 2) nabízí historický přehled vývoje konceptu IQ a inteligence z pohledu asociace mentálních schopností a moci, tedy práva vládnout, již od antiky. Inteligence byla v minulosti i v souvislosti s pseudovědovědami jako eugenika,[6] zneužívána k ospravedlnění některých aktů proti lidskosti, přesněji otroctví a kolonialismus.

Nadřazenost bílého muže vůči domorodým obyvatelům a původním kulturám. Arogance, která dávala „inteligentním" právo ničit ty „podřadné". Inteligence je v tomto pojetí zneužita ve smyslu ideologie, která, jak argumentuje Cave (2020, s. 6), ale možná jednou povede k paradoxní situaci, kdy člověk stvoří umělou superinteligenci. Právě z historické zkušenosti zneužití inteligence vůči „podřadným" a „méněcenným" pravděpodobně pramení obava některých

[6] O objektivitě statistiky, o jejích kořenech, vzniku a původních cílech zakladatelů moderních statistických metod pojednává článek Aubrey Clayton (2020). Statistické metody měly původně podpořit a vědecky podložit předpoklady jejich zakladatelů a zároveň představitelů eugeniky. Původně vědecký směr, zneužitý k podpoře kolonialismu, otroctví a následně i holokaustu za druhé světové války, měl prakticky za cíl prokázat nadřazenost bílých Evropanů nad etniky z jiných kontinentů, které si západní civilizace podmanily.

vědců[7], že tato superingeligence, pokud kdy nastane, nezbytně zničí lidstvo, protože se též stane podřadným a nadbytečným.[8]

Ačkoliv tento stav vývoje UI nemusí nastat, je v kontextu etiky UI zajímavé vzít tuto historickou zátěž výzkumu v oblasti technických věd a kultury některých oborů v úvahu právě v souvislosti se snahou tuto kulturu měnit ve prospěch diskuse o etických otázkách spojených s výzkumem, v tomto případě konkrétně UI (Cave, 2020, s. 5).[9]

1.1.2 Umělá inteligence

Počítačová/„umělá" inteligence se přirovnává k té lidské. Hlavními milníky, jak její pokrok změřit a ověřit, bylo postupné zvládnutí kognitivních úkolů aspoň na takové úrovni, jak by to zvládl člověk. Příklady těchto historických milníků vývoje UI jsou takové činnosti, které byly tradičně považovány za velmi složité pro člověka, které by vyžadovaly mimořádný intelekt. Příkladem může být schopnost hrát šachy nebo hru Go, řešit komplexní matematické úlohy a teorémy, zpracovávat informace, ale i porozumění přirozenému jazyku, překlady, kreativita a umění nebo schopnost komunikovat jako člověk tak, aby ani člověk sám nepoznal, že jedná se strojem.

Poslední zmiňovaný příklad je tzv. Turingův test. Alan Turing sám jej nazýval „imitační hrou". Jak moc je UI schopna napodobit člověka, že to ani

---

[7] Coeckelbergh (2020, s. 13) zmiňuje např. Kurzweila, Bostroma, Hartariho a Tegmarka. Cave (2020) uvádí, že Max Tegmark je ředitelem Future of Life Institute. 22. 3. 2023 publikoval výzvu k zastavení vývoje umělé inteligence (UI) na 6 měsíců a celkovému přehodnocení etických otázek a možných rizik spojených s vývojem UI. Zdánlivá dokonalost odpovědí ChatGPT-4 od firmy OpenAI, tehdy vyvolala paniku u některých firem, významných technologických lídrů, jako např. Elon Musk (Tesla, SpaceX, X – dříve Twitter) a Steve Wozniak (z firmy Apple), kteří pod tuto výzvu připojili své podpisy. K reálnému zastavení vývoje však nedošlo. Zajímavým pozitivním efektem však byla celospolečenská diskuse o nebezpečí a etických otázkách spojených s vývojem UI.

[8] Vznik superinteligence jako jednoho z možných existenciálních rizik je popsáno v podkapitole 1.2.

[9] Konkrétní argumenty podporující uvedené tvrzení se nacházejí v podkapitole 3.2 s názvem The Fetishization of Intelligence, and Diversity in the Technology Sector, kde Cave cituje i Cynthii Lee (2017).

člověk sám nepozná? Diskuse odborníků z přelomu let 2022 a 2023 naznačují, že v podstatě veškeré tyto milníky již byly postupně překonány, zvlášť s ohledem na nejnovější vývoj v oblasti komunikace a psaní textů prostřednictvím tzv. ChatuGPT. Tato technologie je založena na velkých jazykových modelech (LLM).

Termín „umělá inteligence“ (dále zkr. UI) vymyslel a spojil počítačový vědec John McCarthy, když v létě roku 1956 uspořádal workshop, který se nakonec stal průlomovou událostí a znamenal počátek historie tohoto oboru. Tento workshop, s názvem „Dartmouthský letní výzkumný projekt o umělé inteligenci“, měl ambiciózní cíl a svým stejně pokrokovým názvem měl přilákat pozornost špičkových odborníků, aby pomohli prozkoumat a nalézt odpověď na otázku, kterou v roce 1950 vznesl anglický matematik Alan Turing ve svém vědeckém článku „Výpočetní technika a inteligence“, publikovaném ve čtvrtletníku Mysl: „Mohou stroje myslet?“[10] (1950, s. 433-460; Hosanagar, 2019, Christian, 2020, Coeckelbergh, 2020 a další).[11]

„UI je věda o systémech imitujících chování inteligentních živočichů (Řehořek & Surynek, 2023)." Podle autorů Bartneck et al (2021, s. 8) je definice UI zatím volatilní, protože se tento obor dynamicky vyvíjí. Další z možných definic UI je: „Schopnost systému správně interpretovat externí data, učit se z takovýchto dat a toto poznání využít k dosažení specifických cílů a úkolů prostřednictvím flexibilní adaptace (Kaplan and Haenlein in Bartneck et al., 2021, s. 8). Poole a Mackworth (Ibid.)[12] definují UI jako „obor, který studuje syntézu a analýzu výpočetních agentů, kteří jednají inteligentně“. To nastává pokud tento „agent jedná přiměřeně v souladu s jeho situací a cíli, je dostatčně flexibilní, aby se přizpůsobil měnícímu se prostředí a změně cílů, je schopen učit se ze

[10] A. M. Turing (Oct., 1950), Computing Machinery and Intelligence; Mind, New Series, Vol. 59, No. 236 pp. 433-460; Oxford University Press on behalf of the Mind Association; http://www.jstor.org/stable/2251299; (accessed 11. 8. 2023).

[11] Jedná se o filosofickou otázku, na kterou se snažili najít odpověď již v 17. století filosofové jako „Descartes, Pascal, Hobbes či o století později La Metrie. Tyto úvahy neposkytovaly [zatím] žádný návod, jak [technicky] strojového myšlení dosáhnout“ (Mařík et. al, 1993, 15).

[12] Vlastní překlad autorky.

zkušenosti, a je schopen učinit přiměřená rozhodnutí s ohledem na svá výpočetní a percepční omezení.“

Virginia Dignum při popisu a klasifikaci UI cituje Russella a Norviga (2019, s. 12). Zde uvádím pro představu, jak komplexní tento pojem je a těchto klasifikací a jejich metod je ještě více.[13] Inteligentní systémy tito autoři rozdělují do následujících tříd:

- Systémy, které myslí a uvažují jako lidé – se zaměřením na kognitivní modelování – tzv. kognitivní architektury a neurální sítě
- Systémy, které jednají jako lidé – se zaměřením na simulaci lidské aktivity, jejíž úspěšnost se posuzuje provedení testů vycházejících z přístupu Turinga (původní Turingův test – viz. text výše v této podkapitole).
- Systémy, které racionálně uvažují za použití – ty využívají přístupy založené na logice, aby modelovaly nejistotu a byly schopny řešit komplexitu (jsou to tzv. řešitelé problémů – problem solvers, inference, prokazatelé teorémů a optimizace).
- Systémy, které jednají racionálně - se zaměřením na agenty, kteří maximalizují očekávanou hodnotu jejich výkonu v rámci daného prostředí.[14]

John Searle rozděloval UI na tzv. slabou (weak), tu která je schopna vykonávat jen úzce specifikovaný úkol. Většina moderních systémů UI spadá do této kateorie. Opakem je tzv. silná UI (strong), která by se blížila lidským kognitivním schopnostem. Ta zatím, podle vědců neexistuje (Bartneck et al.,

[13] Těchto metod je více. Virginia Dignum dále (2019, s. 12) uvádí jako vhodný zdroj ke studiu UI knihu: Domingos, P. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, 2015.

[14] 6. 3. 2024 vydala OECD revidovanou verzi definice UI systému: Je založený na mechanickém fungování a pro explicitní či implicitní cíle vyvozuje ze vstupů, které obdržel, jak vygenerovat výstupy v podobě předpovědí, obsahu, doporučení či rozhodnutí, která mohou ovlivňovat fyzické nebo virtuální prostředí. Různé UI systémy se od sebe po nasazení liší v úrovni autonomie a adaptability (OECD.AI, 2024). Vlastní překlad autorky.

2021, s. 10), realita se však blíží systémům, které jsou schopny řešit vícero problémů najednou, které nemusejí být úzce specifikovány, jako v případě slabé UI.

UI je souhrnný termín, kam patří vědní disciplíny jako: zpracování přirozeného jazyka (NLP), zpracování velkého objemu dat (big data), robotika, multiagentní systémy, teorie her, strojové vnímání a rozpoznávání (výstupem je např. automatické rozpoznávání obličejů – face recognition), strojové učení[15], nebo softcomputing a biologicky inspirované algoritmy.

Jedná se tedy o vědecký obor, ale termín UI se používá i ve významu nástroje v celé řadě aplikací. Může být totiž aplikována v nejrůznějších oblastech od logistiky (např. řešení optimalizovaného nakládání kontejnerových lodí), autonomní automobilové přepravy, v doporučovacích systémech, zpracování nestrukturovaného textu, algoritmickém obchodování či na burze s cennými papíry.

Aplikuje se také do nejrůznějších objektů, strojů a robotů, které se pohybují po povrchu či létají a vykonávají různé úkoly v oblasti logistiky, zdravotnictví, osobních asistenčních či rehabilitačních robotů, až po vyžití v armádě. Tomu se říká ztělesnění UI, anglicky embodyment. Jedná se o velmi úzce specifikované úkoly a robot, který pracuje v logistickém centru není zároveň schopen hrát šachy, vyklízet myčku, pohybovat se v náročném terénu a chatovat. Je zároveň závislý na výdrži baterií a odolnosti materiálu, ze kterého je vyroben. Díky tomu je v současné době jeho autonomie velmi limitována.[16] A z mého rozhovoru s MH

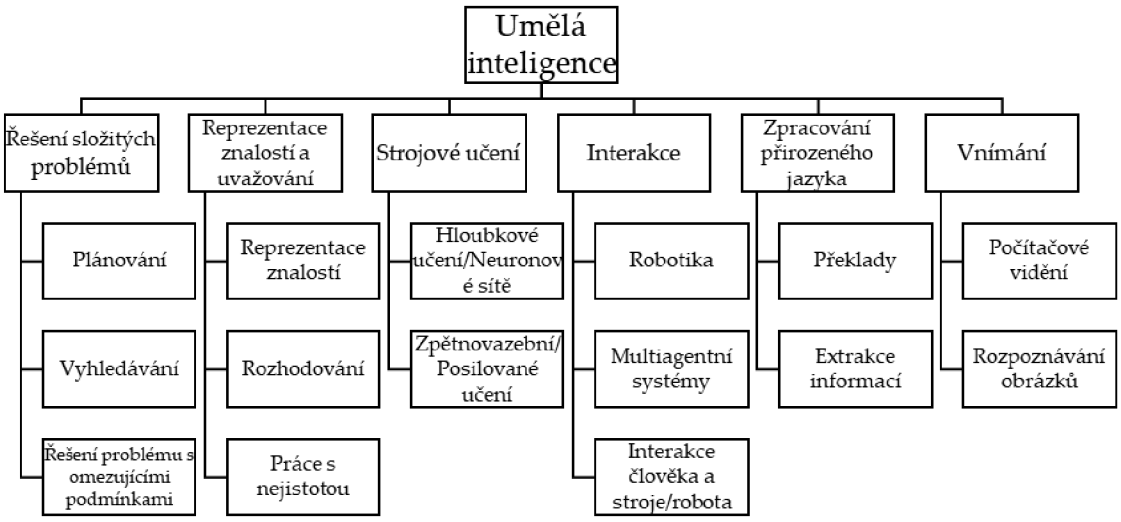
---

[15] Strojové učení lze kvalitně natrénovat pouze pomocí kvalitních vstupních dat. Kvalitu dat ovlivňují různé faktory, mezi něž patří dostupnost těchto dat, správné roztřídění/kategorizace a spolehlivost jejich zdrojů. S dostupností dat souvisí jejich ochrana (např. GDPR omezuje využití osobních dat, což je špatné pro strojové učení, ale pochopitelně důležité pro entity, kterým tato data patří, např. občané, pacienti, minority, firmy atd.). Je zde rozpor mezi využitelností dat (usability) a bezpečností (security) (Holzinger et. al., 2018 s. 5).

[16] MH, rozhovor 16. 11. 2022: „No tak já se vždycky snažím uklidnit tu veřejnost, že opravdu to jako není jako zítra, prostě že my se nebojíme, že ten iCUB (pozn. humanoidní robot) vykráčí z té laboratoře se svým vědomím a bude chtít ovládnout svět, že to je prostě úplně jako nesmyslná úvaha. Že to prostě tak není protože není žádný důvod, se něčeho takového bát, protože ty

ze 16. 11. 2022 vyplývá, že humanoidní robot, který plně ve všech funkcích nahradí člověka, je zatím pouze představa spíše ze sci-fi literatury (iDNES.Cz, 2022, October 21).

Názornější je následující graf – obrázek č. 2[17], ze kterého, jak Dignum uvádí (2019, s. 12), je patrné, že jednotlivé podobory jako např. strojové učení (Machine Learning) a zpracování přirozeného jazyka (NLP) spolu nemají mnoho



Obrázek 2: Ilustrační diagam členění jednotlivých oborů spadajících pod UI (podle Dignum, 2019, s. 12).

společného. Jedná se pouze o ilustrační model, který, jak Dignum sama uvádí, není v rámci vědecké komunity v oblasti UI nijak standardizován či všeobecně přijímán. Členění zároveň může být mnohem podrobnější a nové podobory stále vznikají.

V robotice bychom mohli dále pokračovat ve výčtu různých podoborů jako např. industriální/průmyslová, humanoidní či sociální robotika, která zkoumá oblasti, kde může robot[18] člověku pomoci nebo ho v některých hostilních prostředích – od nemocnic v době pandemie, po průzkum mořského dna, či vesmírných těles - nahradit, doplnit nebo posílit lidské dovednosti a schopnosti

technologie jsou prostě hodně primitivní a asi důležitý jako říct, že ti roboti dělají primárně to, k čemu je naprogramujeme. A vlastně žádný roboti nedělají něco, co sami chtějí."

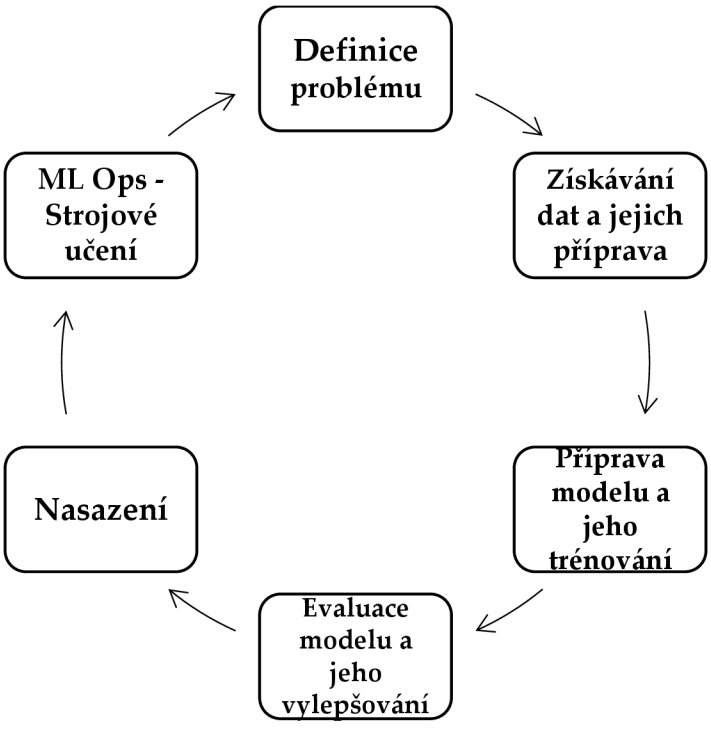
[17] Vlastní překlad autorky.

[18] V průmyslové výrobě se často jedná o kolaborativní roboty – tzv. koboty (cobots).

(tzv. augmentace). Zabývá se komunikací mezi člověkem a robotem – verbální či neverbální, vnímáním robotů lidmi v různém prostředí od domácností po veřejné instituce, tím, jak budovat důvěru člověka a důvěryhodnost robotů, i toho, jak robot může ovlivnit lidské chování, negativně i pozitivně (tzv. nudging).

S tím vším je spojeno mnoho etických otázek, které tento multidisciplinární obor také zkoumá, např. budování přátelského pouta mezi člověkem (seniorem, dítětem) a humanoidním robotem, antromorfizace, sběr a přenášení dat robotem o jeho uživatelích/pozici 3. straně – např. firmě, výrobci, ovlivňování chování, vliv na psychiku, atd.

Ke zvážení možných etických dopadů a rizik spojených s výzkumem a vývojem UI je potřeba dále v textu zmínit i jednotlivé fáze, které ve výzkumu probíhají (EC DG RI, 2022) a druhy výzkumu.



Obrázek 3– Životní cyklus UI (Datascience)

1.1.3 **Základní (badatelský) výzkum v UI**

Je teoretický - jedná se např. o oblast plánování, teorie her, kam patří i kooperativní hry, teorie algoritmů, kam spadá např. problematika obchodního cestujícího (tzv. watchman route problem), který se pohybuje za určitých více či méně komplexních podmínek v určitém regionu, v klasickém „euklidovském prostoru“[19], nebo prostoru s jakýmkoliv jiným tvarem - např. sféry, krychle či množiny „konvexních elipsoidů a polyhedronů“ (JD, 2022, 16. 11.).[20] GNG (Growing Neural Gas), která se využívá při plánování bezpečnostních koridorů pro přistávání letadel. Řeší tedy velmi teoretické, matematické problémy.

Podstatou základního výzkumu je podle Drozenové (2010, s. 281) ...

„...sledovat dlouhodobé a perspektivní cíle a jeho výstupem je nové poznání na určité úrovni obecnosti, které však vytváří předpoklady pro nové možnosti praktických efektů.“

Výsledkem základního výzkumu jsou, podle Zákona č. 130/2002 Sb, odst. k) 1. ...

„…nové vědomosti o základních principech jevů, procesů nebo pozorovatelných skutečností, které jsou publikovány podle zvyklostí v daném vědním oboru“ (Česko, 2002).

Podle Článku (227) materiálu OECD DSTI/EAS/STP/NESTI (93) 6 lze základní výzkum dále rozdělit na: ...

„...čistý základní výzkum neboli badatelský výzkum, který je prováděn v zájmu rozvoje poznání, a to bez úsilí o hospodářský či sociální přínos (ani dlouhodobě) a také bez

[19] Euklidovský prostor je nejvíce podobný přirozené představě člověka o prostoru, ve kterém se pohybuje.
[20] Vlastní rozhovor, JD, 16. 11. 2022. Doktorand/-ka v AIC, Katedry počítačů, FEL ČVUT.

snahy o aplikaci výsledků na řešení praktických poměrů, i bez snahy o předání výsledků těm, kteří jsou za využívání vědeckých poznatků odpovědni; orientovaný základní výzkum, který je prováděn s očekáváním, že vytvoří širokou bázi poznatků, která pravděpodobně bude základem pro řešení již rozpoznaných či předpokládaných (aktuálních či budoucích) problémů, či objevujících se možností využití."

Cílem je tedy posunout hranice vědeckého poznání a můžeme si položit otázku, zda toto vědění je automaticky morálně/eticky neutrální, případně, jaké možné etické souvislosti se základním výzkumem v UI mohou souviset. Ve zdravotnickém výzkumu či bioetice by se i v rámci základního výzkumu vyžadovalo posouzení a povolení výzkumu etickou komisí (Drozenová, 2010; Veselská, 2023; Marušáková, 2023). Jedná se tam o minimalizaci negativního dopadu výzkumu na člověka, případně na laboratorní zvířata, obsahuje to oblast nakládání s pacientskými osobními daty, která je ošetřena GDPR, povolenými metodami výzkumu atd.

V oblasti UI by analogicky též měla podobná opatření probíhat. Zároveň většina vstupních dat jsou ve formě open source nebo jsou poskytnuta třetí stranou. Univerzita většinou (kromě např. ÚFAL na Matematicko-fyzikální fakultě Univerzity Karlovy, která vytváří své vlastní velké jazykové modely, jazykový korpus) nevytváří ani nevlastní datasety, které pro výzkum využívá (rozhovor, LK, 2024). Jaké tedy má možnosti kontrolovat jejich kvalitu a reprezentativnost, případně jejich zdroj a další podrobnosti.[21]

Etické aspekty UI lze řešit i přímo technicky nebo se podobné cesty zvažují, což je často předmětem nebo součástí výzkumu samotného jako v případě trasparentních/vysvětlitelných algoritmů (např. jako v projektu AutoFair financovaném v rámci Horizon Europe, https://humancompatible.org). Dalším směrem je zjišťování možností, jak technicky zakódovat morální zásady a

[21] Zformulováno na základě rozhovor s LK na téma odpovědnost vědců v oblasti UI a zkušenosti se spoluprací s univerzitní etickou komisí. Osobní rozhovor proběhl 13. 3. 2024 na Katedře počítačů, ČVUT FEL.

hodnoty, které by byly v souladu s těmi lidskými, přímo do algoritmů (AI alignment).

1.1.4 Aplikovaný výzkum

přináší podle definice Zákona č. 130/2002 Sb, odst. k) 2.: ...

„...nové poznatky a dovednosti pro vývoj výrobků, postupů nebo služeb, poznatky a dovednosti uplatněné jako výsledky, které jsou chráněny podle zákonů upravujících ochranu výsledků autorské, vynálezecké nebo obdobné činnosti nebo využívané odbornou veřejností či jinými uživateli, nebo poznatky a dovednosti pro potřeby poskytovatele, využité v jeho činnosti, pokud vznikly při plnění veřejné zakázky, nebo ve vývoji návrhy nových nebo podstatně zdokonalených výrobků, postupů nebo služeb.“

Aplikovaný výzkum lze podle Článku (232) materiálu OECD DSTI/EAS/STP/NESTI (93) 6 dále rozdělit na: ...

„...všeobecný aplikovaný výzkum, který je soustavným zkoumáním za účelem získání nových poznatků, které ještě nedosáhlo stadia s jasnou specifikací cílů pro jeho aplikace; specifický aplikovaný výzkum, který je rovněž soustavným zkoumáním za účelem získávání nových poznatků, ale směrovaných k specifickému praktickému cíli s jasnou aplikací výsledků“ (Úřad vlády ČR).

1.1.5 Interdisciplinarita

Jan Kulveit[22] v přednášce „Multidisciplinarita ve vědě“ (Czexpats in Science, 2021) vysvětluje, kdy je interdisciplinarita nutná. Rozlišuje, mezi interdisciplinaritou skutečnou a umělou. Ta skutečná, podle jeho názoru, vychází

[22] Jan Kulveit se zabývá existenciálními riziky spojenými s výzkumem a vývojem UI (mimo další komplexní problémy ohrožující lidstvo). Založil v rámci Centra pro teoretická studia, které je společným pracovištěm Univerzity Karlovy a Akademie věd vlastní výzkumnou skupinu The Alignment of Complex Systems (ACS).

z dobré výzkumné otázky, na kterou nedokáže odpovědět jeden obor. „Odpověď vyžaduje kombinaci metod, znalostí anebo přístupu z více oborů“, případně: „Odpovědět by dokázal obor, který by mohl existovat, ale zatím neexistuje.“

Problémy s interdisciplinaritou ze zkušeností J. Kulveita nastávají např. v recenzním řízení, kdy odborníci s určitou specializací nemají stejný mezioborový přesah jako hodnocený vědecký článek/projektový záměr. Dále naznačuje, že incentivy k mezioborovému výzkumu ze strany grantových výzev / poskytovatelů nefungují tak efektivně, jak by mohly. Účastníci přednášky se shodli na tom, že současné pojetí vědy, tak jak jsou jednotlivé vědní disciplíny roztříděné do jednotlivých oborů, není zcela efektivní při potřebě vědy řešit stále komplexnější výzvy.

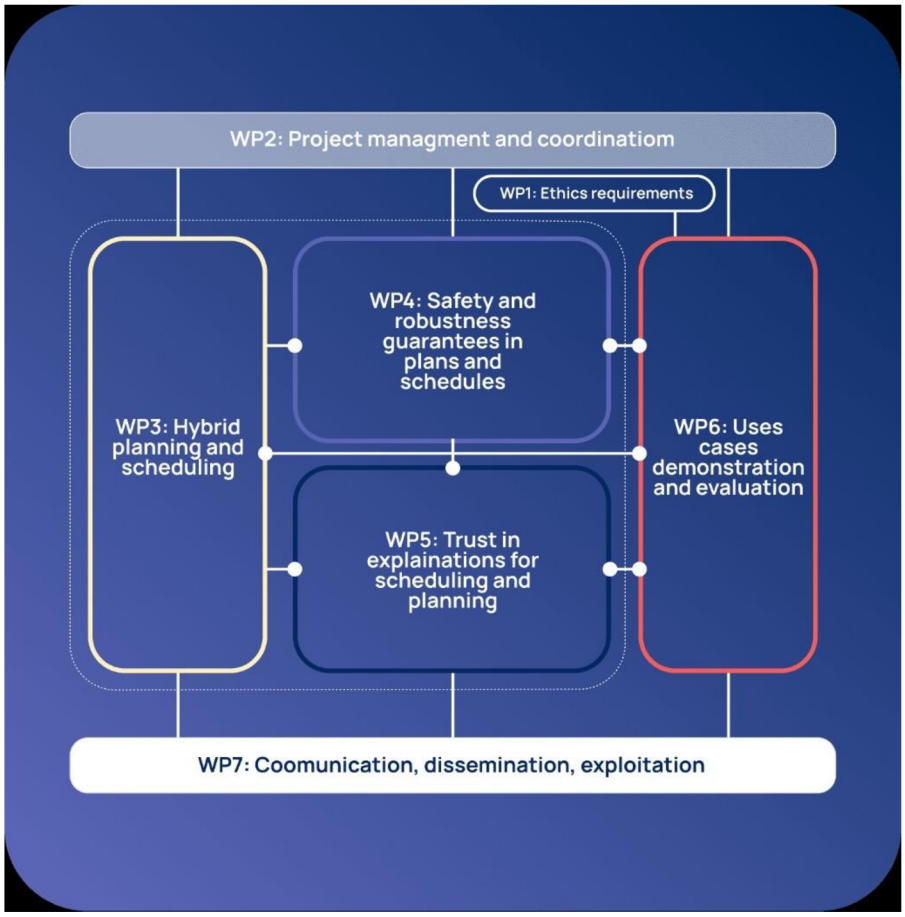
A dále byla diskutována i schopnost vědců komunikovat napříč obory, kdy J. Kulveit zdůraznil, že bez snahy naučit se aspoň částečně ostatním oborům, tedy neustále se vzdělávat, není tato komunikace efektivní. Vědci s hlubokým vhledem do určité problematiky nemají automaticky schopnost porozumět jinému oboru. U příbuzných oborů zase hrozí, že je stejná terminologie využívána v jiném kontextu, což je matoucí a opět neefektivní.

Prof. Veselská (dále v 1.1.3., 1.1.5 a kapitole 3), „předsedkyně Etické komise pro výzkum MU, členka Bioetické komise Rady pro výzkum, vývoj a inovace a Etické komise Ministerstva zdravotnictví ČR. V roce 2022 se stala jako vůbec první zástupce ČR členkou Evropské skupiny pro etiku ve vědě a nových technologiích (EGE).“ Prof. Veselská je přesvědčena, že **etika výzkumu je z podstaty mezioborová záležitost** (Marušáková, 2023). Pro hodnocení etiky daného výzkumu/plánovaného výzkumného projektu doporučuje celouniverzitní (ne fakultní) Etické **vědecké** komise[23] z důvodu, že jsou mezioborové. Sama využívá v praxi poradenství nejen etiků, ale ještě i právníků, ale výzkum může být veden různým způsobem a jak říká: „V ČR není

[23] K otázce etických komisí ještě v podkapitole 3.2.3. Výzkumné instituce.

systematicky ošetřeno dodržování etických standardů v psychologickém, behaviorálním či jiném typu výzkumu," což je relevantní i v případě UI a jejím možným účinkům nejen na člověka, ale na celou společnost.

Aplikace UI je prakticky ve všech sférách lidské činnosti. Jak vyplynulo z rozhovorů s vědci UI z kateder počítačů a kybernetiky FEL ČVUT, pokud se jedná o aplikovaný výzkum, interakce s jinými obory je zde pravidlem. Pokud se jedná o humanoidní robotiku, je nezbytná znalost kognitivní a vývojové psychologie, neurověd, biologie člověka (kromě vědců, kteří se přímo zabývají oborem bioinformatika), ale i jiných živých organismů, např. hmyzu. Využívání technologií jako rozpoznávání obličejů v jiném kontextu. Tato technologie po natrénování pomocí strojového učení z tisíců fotografií rozpoznává např. z unikátních znaků na krunýři individuální jedince, jejich pohyb v rámci teritoria a je tak možné sledovat výskyt např. na konkrétních středomořských ostrovech a migraci ohrožených druhů zvířat.



Obrázek 4– „Pracovní balíček – WP – zabývající se etickými požadavky v případových studiích projektu „Tuples – trustworthy AI“.[1]

Robotika také vyžaduje vhled a znalost prostředí, kde budou řešení aplikována, např. využívání dron v kontextu památkové péče, výpočtu rozsahu požáru, plánování v logistice nebo autonomní přepravě, optimalizace v energetice, využití robotů při vojenských nebo záchranných operacích v různém terénu či pod zemí, či jako hasicí roboty. Jedná se potom o znalost i praxe a kontextu, v jakém budou UI řešení či roboti nasazeni, příslušné legislativy, stavu poznání a jednotlivých aktérů z jiných oborů a profesí, se kterými spolupracují. Inter/multidisciplinalita je tedy pro vědce UI nezbytná a běžná.

Na Obr. 3 je vidět zapojení pracovního balíčku „Etické požadavky“ do organizační struktury projektu. Vzhledem k záměru projektu vytvořit UI řešení v kontextu konkrétních případových studií tak, aby splňovalo princip transparentnosti, robustnosti a bezpečnosti, jsou technická řešení konzultována

s odborníky na organizační psychologii a další humanitní obory. V oblasti etiky není ještě vybudována „kultura“, zvyk, že je potřeba uvažovat i o dopadech svého výzkumu v různém kontextu, nejen tedy o dosažení cílů, ale přemýšlet v horizontu efektu těchto výsledků na člověka, potažmo celou společnost (to záleží samozřejmě na oboru UI, zvolené metodě, praxe, kde bude UI řešení aplikováno, atd.). Neexistuje zde řešení, které by pasovalo na všechny případy a i v případě stejného vědce bude efekt jeho výzkumu vždy jiný. Nikdy se tedy nebude jednat o nějakou rutinu.

Pokud se podaří změnit porozumění etické dimenzi výzkumu jako něčeho, co může přispět k ještě lepším výsledkům výzkumu (excelenci) nebo vůbec k lepšímu hodnocení projektového záměru (vyšší úspěšnost – která bude i měřitelná, protože písemné hodnocení projektových návrhů HE bývají poměrně podrobná), je zde šance na lepší a přirozenější integraci etického rozměru do výzkumu UI.

Na druhou stranu to vyžaduje i schopnost filosofů, etiků a sociologů porozumět aspoň do určité míry výzkumnému záměru i po technické stránce a kontextu impaktu (viz pohled J. Kulveita na efektivní interdisciplinaritu výše). Případně, ideálně, zaměřit se na vzdělávání studentů technických oborů v oblasti etiky, aby sami získali kombinované (interdisciplinární) znalosti v kontextu své specializace.

Z výše uvedeného vyplývá, že etika by neměla být chápána jen ve smyslu „odškrtnout“ si nějakou povinnost nebo jako nepříjemná bariéra na cestě ke grantu na výzkum nebo kontrola. Potom bude interdisciplinární spolupráce v zájmu vývoje etické a společensky prospěšné umělé inteligence skutečně reálná.

1.2 Metodologie - jak výzkum probíhal, jednotlivé etapy poznání:

1.2.1 Teoretická část:

- Literární rešerše k tématu: etika UI - výstupem je zjištění, že odborné literatury na téma etiky v UI (etické, zodpovědné, atd.) UI je velké množství, stejně jako konkrétně definovaných principů „zodpovědné UI“, jen to není reflektováno v praxi nebo změně interní kultury organizace ve výzkumu a výuce.
- Potvrzení, z rozhovorů a kontrolou kurikula, že studenti technických oborů na bakalářském i magisterskému stupni studia zabývající se umělou inteligencí, mají pouze kurz filosofie v prvním ročníku. V oboru etiky žádný konkrétní kurz nemají, ani formou např. řešení aktuálních technicko-společenských otázek prostřednictvím mezioborové diskuse (i když se chystá akreditace nového magisterského programu se zohledněním výuky SSH) nebo studenti využívají online kurz Ethics of AI. Zapojením se do realizace projektů Horizon Europe se však studenti na doktorské úrovni studia prakticky seznamují s vlastní výzkumnou praxí, zde je interdisciplinarita běžná. Je to pro ně skvělá praxe.
- Stáž v kanceláři CZELO – Brusel. Rozhovory se zástupci a zástupkyněmi institucí, které tvoří „kontrolní mechanismy“ vědy, vědeckou politiku nebo řeší aktualizace oblastí, které vědecký ekosystém přímo ovlivňuje – zákony („Digital“ Acts, AI Act), reforma systému hodnocení vědy (EUA), etické otázky (podmínky) v projektech Horizon Europe (EC DG RI, I. Karatzas), změny v přístupu k výuce inženýrských oborů (SEFI), posilování spolupráce evropských univerzit formou Univerzitních aliancí.
- Účast na interdisciplinárních a odborných konferencích: Kognice a umělý život 2022, ICSR 2022 (14th International Conference on Social Robotics), EARMA 2023, NCURA + MUNI Grants Week Brno 2022, 2023,

konference v rámci CETE-P k UI etice s M. Coeckelberghem atd. Rozhovory s vědci a vědkyněmi o etice UI, výuce a zkušenostech

- Zjišťování, jak by šlo změnit výuku. Možnosti, které jsem předpokládala: prostřednictvím vzdělávání, diskuse, společných projektů STEM a SSH oborů, interdisciplinaritou samotných vědců. Etika by měla být součástí základního výzkumu – vědecké integrity vědců

1.2.2 Empirická část:

- Teoretická část pokračuje studiem odborné literatury a zdrojů, zároveň začíná empirická část – rozhovory a interakce s vědci, založení pracovní skupiny Responsible AI:
- Kaufmann - Chápající rozhovor – studium metodologie rozhovorů. Probíhají první rozhovory s vědci a doktorandy od listopadu 2022, kde si ověřuji současný stav a jejich přístupu k etice UI. Zároveň je to první zkušenost, kdy se s vědci bavím o jejich práci a vědeckých tématech. Postupně zjišťuji, že etický rozměr výzkumu a umělé inteligence zajímá více vědců/doktorandů, než jsem si původně myslela. Strukturovaných rozhovorů proběhlo celkem 12.

  Poté už probíhaly volnou formou a ve velkém množství pro upřesnění různých detailů nebo aktuální situace v oblasti UI. Tyto pravidelné interakce vedly k založení nezávislé pracovní skupiny Responsible AI (RAI) v rámci Katedry počítačů (ČVUT FEL) – cca 6 pravidelných členů, další jsou připojeni a sledují kanál RAI na Slacku, kde si sdílíme akademické články, informace o akcích a konferencích k tématu etiky UI..
- První schůzka proběhla v červenci 2023 a od té doby se skupina schází pravidelně jednou týdně. Předmětem činnosti jsou diskuse na různá odborná témata k etice UI, která postupně redefinujeme. Skupina též slouží jako reading group pro studenty doktorského studia. Pokud se někdo

zúčastní odborné přednášky nebo konference, sdílí informace s ostatními, networking. Zorganizovali jsme přednášku v rámci AI Days pod záštitou prg.ai 4. 11. 2023 na téma UI etiky, které se zúčastnilo více ne 50 osob z ČVUT i jiných organizací.

- V rámci skupiny RAI se vedly i diskuse na téma potřeb změn ve výuce směrem k profesní etice a etice UI. Diskutovány byly různé možnosti. V rámci RAI skupiny – diskuse na téma vzdělávání a potřeby studentů bakalářského a magisterského studia v této oblasti. V rámci skupiny se všichni členové shodli, že je toto vzdělávání důležité. Na druhou stranu, na možnostech praktické realizace ne.
- Založení knihovničky na katedře se základní aktuální literaturou v oblasti AI ethics, s autory jako Dignum, Boddington, Coeckelbergh, Floridi atd.
- Administrativní převzetí agendy Horizon Europe na katedře, cca 8 projektů. Většina z nich přímo či nepřímo řeší etické otázky spojené s UI ve spojení s různými obory a využitím: biologie, personalistika, finanční sektor, spojení s průmyslem atd. Tyto zkušenosti se také propisují do textu DP. Dalším přínosem je účast na projektových mítincích, kontakt se zahraničními partnery a manažerskými týmy. Ověření, že interdisciplinarita a multidisciplinarita jsou skutečně základem vědecké spolupráce v těchto projektech, které řeší kritické společenské problémy.
- Studium aktuální literatury ve vztahu k „mechanismům“ – strategické dokumenty na podporu VaVaI UI, aktéři v rámci českého a evropského vědeckého ekosystému. Jedná se o komplexní systém, který se neustále vyvíjí.
- Využití profesní sítě LinkedIn k přímému propojení s vědci a vědkyněmi v oblasti etiky UI, research managementu, interdisciplinárních výzkumných center etiky UI, univerzity a další organizace. Sledování konferencí a aktualit.

- Zpracování konečné verze DP. Činnost skupiny RAI pokračuje i nadále.

1.3 Popis současných výzev spojených s výzkumem, vývojem a aplikací UI

„Digitální sféra dobývá a mění vše, co známe, ještě dřív, než o tom začneme přemýšlet a rozhodovat. Propojený svět vynášíme do nebes za to, jak všemožně rozšiřuje naše schopnosti a možnosti, jenže dal vzniknout i zcela novým oblastem úzkosti, nebezpečí a násilí, a zároveň se z něj vytrácí pocit, že lze předvídat budoucnost"
(Zuboff, 2022, s. 18).

Vývoj umělé inteligence tak, aby prospívala lidstvu.[24] Jedná se, na první pohled, o logický požadavek, na kterém se všechny zainteresované strany prakticky bez zásadní diskuse shodnou. Na druhý pohled je to spíše jen určitá vize, jejíž aplikace bude v praxi velmi složitá. V současné době jsme ve fázi, kdy potřebujeme definovat velmi konkrétní kroky, které by měla vědecká obec, ale především velké korporace vyvíjející aplikace a jinak využívající UI, přijmout v jednotlivých oborech UI, aby to mělo skutečný pozitivní dopad. Problematické je totiž slovo **prospěch**,[25] na to upozorňuje např. Paula Boddington (in Dubber, Pasquale & Das, 2020, s. 132). Dignum (2019, s. 51) uvádí, že „skutečná odpovědnost v oblasti UI se neprojevuje pouze v designu těchto technologií, ale tím, jak si definujeme úspěch."

A jak tento prospěch nebo pozitivní/ negativní dopad technologií na společnosti změřit, abychom mohli stanovit měřitelné cíle, strategii jejich

[24] Také překládáno jako: pro blaho lidstva, v zájmu lidstva, v souladu s lidskými zájmy (angl. AI for the common good and benefit of humanity; human-centered alignment atd.)
[25] Také termín lidstvo (angl. v souvislosti s principy odpovědné UI vyjadřováno slovy humanity, humankind, people) je poměrně příliš široký. Jak definovat, co je v zájmu všech lidí? Když se hovoří např. o společnosti, nejedná se pouze o skupinu/množinu individualit (Sacks, 2021, 12).

dosažení a kontrolovat jejich plnění. Inspirací mohou být cíle udržitelného rozvoje (SDGs – viz následující text) nebo oblast měření společenských změn (social impact) a sociálních inovací, kterými se již kolem 20 let zabývá např. New Economics Foundation ve Velké Británii. Tamtéž také proběhl výzkum CFI,[26] který financovala nadace Nuffield.[27]

Pod pojmem prospěšné využití UI si pravděpodobně každý představí celou řadu praktických aplikací např. v oblasti zdravotnictví, které je v souvislosti s pozitivním využitím UI často zmiňováno, kde by mohlo využití UI skutečně pomoci. Čí názor a úhel pohledu by však měl mít přednost, představíme-li si nejrůznější cíle, které jsou vzájemně v opozici nebo přímo v konfliktu? Je lidstvo schopné se vůbec dohodnout na společných cílech a prioritách? Doina Precup (in Hrivňák, 2023), vedoucí kanadského týmu výzkumníků DeepMind, který vyvíjí umělou inteligenci pro Google, potvrzuje, že je pro ni ...

„… obtížná otázka, s cíli kterého člověka má být umělá inteligence sladěna. Druhou otázkou je, zda cíle, které si konkrétní člověk stanovil, jsou pro něj skutečně dobré a zda jsou dobré pro společnost. Když mluvíme o sladění umělé inteligence a lidí – co vlastně chceme sladit? A kteří lidé by měli systému poskytovat zpětnou vazbu?“

I Precup potvrzuje, že řešení těchto otázek je potřeba hledat ve „spolupráci s dalšími obory, jako jsou společenské vědy, ekonomie a právo.“ Tyto problémy nejsou technické, ale spíše sociální či filosofické (Hrivňák, 2023).

Podle Coeckelbergha (2022, s. 4) společnost často považuje UI za neutrální nástroj. Spolu s Reijersem (2020, s. 3) však ve své knize Narrative and Technology Ethics postupně dokazuje, že „technologie není neutrální. Technologie nás

[26] CFI je zkratka pro Leverhulme centre for the future of intelligence, která se zabývá zkoumáním dopadu a souvislostem z využíváním UI ve společnosti. http://lcfi.ac.uk/

[27] Nuffield Foundation si klade za cíl „zlepšovat životní šance lidí a identifikovat cesty jak čelit znevýhodnění a nerovnostem ve společnosti řízené digitálními technologiemi.“ https://www.nuffieldfoundation.org/research.

přesvědčuje, učí, vyzývá, omezuje, poškozuje, a tudíž aktivně či pasivně přispívá k ovlivňování našich etických rozhodnutí a k aktivitám, do kterých se zapojujeme."

Podle Korinka (in Dubber, Pasquale & Das, 2020, s. 475) se využití UI ubírá podobným směrem, jako předchozí průmyslové revoluce, kde byla jejich hnací silou nejdříve pára a později elektřina. S obrovským potenciálem změnit společnost, kvalitu a délku lidského života, ekonomiku a sociální struktury ve společnosti, nabízí podobné změny a příležitosti i využití UI ve všech oblastech našeho života. S tím však přicházejí i reálné či potencionální hrozby, které využití této nové technologie přináší a ještě přinese.

Jak Korinek (in Dubber, Pasquale & Das, 2020, s. 475-6) dále poukazuje, jsou tyto změny, aspoň prozatím, zcela pod kontrolou lidí. „Lidé rozhodují, co, kde a jak inovovat." A mají tedy i zodpovědnost nepodléhat tzv. techno-fatalismu, tedy předpokladu, že náš osud je předurčen či přímo zpečetěn kombinací sil technologického pokroku a „neviditelné ruky trhu".

Druhým extrémem, se kterým se v oblasti přístupu k novým technologiím setkáme, je techno-optimismus. Meredith Broussard (2018) dokonce zavedla termín „**technošovinismus**", kterým označuje kulturu techno-optimistů, převážně mužů bílé barvy pleti, kteří „slepě věří v moc digitálních technologií vyřešit široké spektrum sociálních problémů. Techno-šovinisté se zároveň vyznačují nezájmem o zkoumání jejich negativního vlivu. Dále je definuje vysoká míra biasu. Vykazují minimální známky zájmu o sociální témata, což je provázeno jejich vírou v individualismus a disruptivní inovace" (Cobbe, J., 2022, s.3).[28]

Tyto určité **rozpory v kultuře a překážky v komunikaci** a vzájemné spolupráci mezi obory společenskovědními a humanitními na jedné straně a

[28] Příkladem mohou být výroky Yanna LeCuna, šéfa výzkumu UI firmy Meta, který se opakovaně nevybíravě vyjadřuje na sociálních sítích o vědcích sdílejících obavy z budoucího vývoje UI. Cobbe, J. (2022).

technickými/inženýrskými obory, nejsou nic nového. Popsal je ve svém proslovu a knize „Dvě kultury“ J. P. Snow již v roce 1959.

Co rozhodně tyto kultury spojuje, nebo by mělo, je nutnost spolupráce při výzkumu s cílem řešit současné globální výzvy spojené s dalším pokrokem v oblasti výzkumu a vývoje UI. Tedy konkrétní cíle a společný účel, pokud je takto definován a finančně a administrativně podporován na nejvyšší politické úrovni. Zabývá se jimi text v podkapitole 1.3 UI jako politické, bezpečnostní a strategické téma.

Nesoulad mezi technologických pokrokem a schopností společnosti tyto změny absorbovat, využít a vypořádat se s jejich negativními dopady, tu byl vždy. Byl tady také často nesoulad mezi cílovými skupinami, pro které technologický pokrok znamenal zvýšení kvality života, ať už v jakémkoliv ohledu, a těmi, kterým naopak přitížil. Radovan Richta, ve své knize shrnující závěry **mezioborového výzkumu**:[29] Civilizace na rozcestí, společenské a lidské souvislosti vědeckotechnické revoluce, z roku 1969, popisoval podobnou situaci. Ani 55 let od jejího vydání nestačilo k tomu, aby byly otázky, které si v této knize klade, zodpovězeny a vyřešeny. Naopak, přibývají další. Co však stále platí, je jeho důraz na důležitost mezioborové spolupráce ve výzkumu.

O tom, zda bude UI využita v mírovém či vojenském kontextu, pro legální nebo nelegální cíle, rozhodují často její uživatelé, odtud termín **„dual use“** – dvojí využití. Například Bezpečnostní Informační Služba (BIS) ve své Výroční zprávě za rok 2022 varovala před syntetickými médii založenými na UI, které vytvářejí umělý/syntetický obsah v podobě videí, fotografií i audia (hlasů osob) a textu (BIS, 2022, s. 19). Pokud jsou využita v oblasti vzdělávání či zábavního průmyslu,

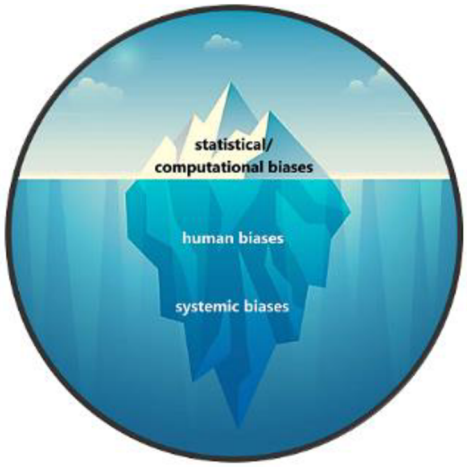
[29] Mezioborový tým byl složen z „pracovníků různých vědních oborů (filosofie, ekonomie, sociologie, psychologie, pedagogiky, politické vědy, historiografie, lékařských věd, teorie architektury a životního prostředí, jakož i technických a přírodovědeckých disciplín aj.) (Richta, 1969, s. 9).

je to v pořádku, ale dvojí využití znamená, že mohou být zneužita i k „páchání zločinů, šíření dezinformací nebo tzv. prank calls.

Jejich negativní dopad se potom násobí „automatizovaným umisťováním do informačního prostoru, např. prostřednictvím sociálních sítí nebo diskusí pod články zpravodajských portálů.“ Potenciální zneužití určitých technických zařízení dvojím způsobem a snahu o jejich vývoz do rizikových (např. válečných) oblastí také v roce 2022 BIS monitorovala.

Výzev spojených s výzkumem a využíváním UI v oblasti jejich vlivu a dopadu na společnost, ekonomiku, zaměstnanost, soukromí a kvalitu života obyvatel, je mnoho. Britský parlament ve své „9. průběžné zprávě ke stavu vývoje regulace, dohledu a řízení v oblasti UI (HC 1769, 2023, 19. 7.)“[30] definoval celkem 12 největších výzev nebo přímo dilemat:

- **Bias** – zkreslení způsobené především špatnou kvalitou nebo zvolenou reprezentativností dat, bias zakomponovaný přímo do algoritmů, ať už



Obrázek 5- Bias jako ledovec, vidět je jen malá část (Schwartz et al., NIST, 2022, i/77).

[30] Zde je možné přímo sledovat záznam slyšení Výboru pro Vědu a technologie Britského parlamentu, kde se vyjadřují ke stavu, hrozbám a vývoji UI zástupci firem Microsoft UK, BT Group nebo Google a profesoři/-rky předních britských výzkumných center v oblasti informatiky: https://parliamentlive.tv/Event/Index/25bb5726-4484-4237-a866-1fcd5ef400a8 Transcript je k dispozici zde: https://committees.parliament.uk/oralevidence/12709/pdf/. (HC 945, 2023. 22. 2.).

vědomě či záměrně a bias jejich developerů (Adrian Joseph z firmy British Telecom, BT Group in HC 945, 2023, 22. 2., s. 13). Joseph uvedl, že první dva případy jsou technicky zřejmé. Pokud je nekvalitní datový input, algoritmy způsobují nekvalitní output. S individuálním biasem inženýrů a techniků je potřeba aktivně pracovat. Jako pozitivní příklad změny uvádí proporcionálnější zastoupení cílových skupin v technických, výzkumných a manažerských týmech. Nejen lepší genderovou, ale i etnickou proporcionalitu. A další opatření v manažerských strukturách firmy, o které se v BT snaží – viz odkazy.[31,32]

Právo na soukromí (Privacy) – vztahuje se k základním lidským právům na autonomii, důstojnost a identitu. Ohrožení vylývají z toho, že technologie založené na UI umožňují a zrychlují sběr, ukládání a šíření/přeprodej osobních informací a jejich využití v různém kontextu bez vědomí a souhlasu dané osoby (např. identity theft, sběr biometrických dat, automatické rozpoznávání obličeje ve veřejném prostoru). Je třea hledat neustálou rovnováhu mezi ochranou soukromí a potencionálním přínosem nasazení UI modelů a nástrojů v určitých oblastech (use cases), jako například při vymáhání práva (HC 1769, 2023, 19. 7., s. 15, bod 49).

- **Zkreslení (Misrepresentation)** – může vzniknout omylem, ale i záměrně jako manipulace s hlasovým záznamem, když voláme třeba do call centra,

[31] **Bias** v datech z pohledu lidských práv je „neoprávněná či nevyvážená přítomnost tzv. „chráněné hodnoty“ v datech, tj. nesoucí informaci o pohlaví, rase, barvě pleti, náboženství, politické příslušnosti apod. Došlo by tak k protiprávní diskriminaci. […] Může se jednat jak o bias ve vývojových datech používaných například v procesu strojového učení, tak v datech používaných pro verifikaci, validaci a kvalifikaci systému.“ Přítomnost této tzv. chráněné hodnoty může být někdy oprávněná, je však potřeba tuto skutečnost prověřit a zajistit reprezentativnost a úplnost dat nesoucí tuto chráněnou hodnotu (Šmuclerová M., Král, L., Drchal, J. et al., 2023c, s. 3).

[32] **BT Group** (2024) definovala 4 principy pro zodpovědné technologie, které pojmenovala **For Good** (s dobrými úmysly), **Accountable** (bereme na sebe zodpovědnost), **Fair** (všichni mají právo na rovný přístup a respekt), **Open** (transparentnost).

při řešení škody s pojišťovnou, za účelem reklamace nebo jiných čistě privátních záležitostí. Hlasový záznam může být prostřednictvím UI zpracován a vygenerovat tak zcela smyšlenou nahrávku s cílem poškodit volajícího/obět, její reputaci. Dále může být použit ke kriminální činnosti vydírání, špionáže, v předvolebních kampaních a manipulaci veřejného mínění, apod.

Při využívání zaintegrovaných funkcionalit nástrojů typu ChatGPT přímo do vyhledávače (Bing) nejsou dotazy zakódovány (encrypted) (Hugh Milward, Microsoft UK in HC 945, 2923, 22. 2., Q137). Tyto informace/záznamy o jednotlivých vyhledáváních jsou potom přístupné jako data o uživatelích k dalšímu zpracování, atd.

- **Přístup k datům** – kvalitní data jsou považována za novou měnu současné ekonomiky, hovoří se o tzv. datovém kapitalismu (West, 2017, s. 20).[33]

  Kvalitní trénovací data představují konkureční výhodu pro výzkumné týmy a firmy (HC 1769, 2023, 19. 7., s. 18). Zároveň jejich vlastnictví představuje obrovskou moc a velkou disbalanci moci vzhledem k tomu, že je většinou vlastní velké korporace, jak tato zpráva uvádí autority v oblasti ochrany spotřebitelů a tržní regulaci.[34] Proto politika EU v oblasti Open Access a sdílení dat je důležitou protiváhou těchto mocných tržních sil – viz strategické plány a dokument Strategická

[33] **Datový kapitalismus** je komodifikace soukromých dat, která způsobuje asymetrické rozložení moci směrem k entitám/firmám, které jsou schopné tato data „těžit". To znamená shromažďovat a využívat a případně zneužívat proti jejich původním majitelům/zdrojům, ať už jsou to občané, zákazníci, voliči, jakékoliv minority atd.(West, 2017, s. 20).

[34] Institut Ady Lovelace (The Ada Lovelace Institute) navrhl legislativní řešení tohoto problému, který by zavázal Big Tech korporace ke zpřístupnění jejich datových úložišť pro vědecké účely. Tento přístup obhajují i Creative Commons (s. 18). Paradoxem v tomto případě je, že představitelé těchto korporací si stěžují na plánovanou legislativu v oblasti UI v EU, že zadusí inovace, ale přitom na tom sami mají podíl vzhledem k vlastnictví, moci a prostředkům shromažďovat tyto datasety. (pozn. autorky)

agenda pro výzkum a inovace (SRIA) Evropského Cloudu pro open science, pro posílení konkurenceschopnosti VaVaI EU.

- Přístup ke konkurenceschopnému **výpočetnímu výkonu**, superpočítačům, výpočetním klastrům a kvantovým počítačům s adekvátní možností využívat cloudové služby k ukládání velkého objemu dat. Opět zde existuje nerovnováha mezi nejsilnějšími hráči na trhu (a zároveň vlastníky těchto cloudových řešení/služeb) a výzkumnými centry na univerzitách a ve výzkumných institucích (HC 1769, 2023, 19. 7., s. 18).
- Výzva spojená s tzv. **fenoménem černé skříňky** (black boxu). Tento požadavek souvisí s principy odpovědné/ etické/ transparentní/ vysvětlitelné UI (tzv. XAI, RAI) (viz kapitola 2 Etika UI). Tento problém vzniká u systémů založených na technikách strojového učení – příkladem jsou neuronové sítě. „Transparentnost je třeba odlišit od algoritmické vysvětlitelnosti. Ta se týká způsobu algoritmické interpretace znalostí, tj. vysvětlitelnosti algoritmického zápisu znalostí. Pro uživatele může být i algoritmicky vysvětlitelný systém netransparentní, pokud neposkytuje informace, na základě čeho se rozhodl“ (Šmuclerová M., Král, L., Drchal, J. et al., 2023c, s. 4, 11).

Institut Ady Lovelace (ALI, 2020, 29. 4., s. 3) představil metodiku a nástroje (toolbox) pro hodnocení algoritmických sytémů a jejich impaktu/dopadu. Tento dokument v úvodu vysvětluje základní termíny přístupů k hodnocení algoritmů a jejich společenského dopadu (souvislostí) s tím, že v rámci různých vědních disciplín (informatika, sociologie), v kontextu státní správy a z pohledu veřejnosti, jsou tyto termíny vnímány trochu odlišně.[35]

[35] Tento fenomén bývá častou příčinou vzájemného neporozumění si vědců a zástupců dalších zainteresovaných skupin a partnerů v interdisciplinárních projektech. Klarifikace základních termínů a slovníku ihned na začátku spolupráce je důležitým základem. Vše se ještě komplikuje

Z tabulky v této metodice (ALI, 2020, 29. 4., s. 5) je zřejmé, kdo je za kontrolu zodpovědný a v jaké fázi výzkumu, vývoje, ověřování či nasazení algoritmu by měla proběhnout. Zároveň je zde nezbytná mezioborová spolupráce, protože ověřování probíhá v kontextu různých oborů (finance, sociologie, lékařství, žurnalistika, sociální správa – příspěvky na péči, životní prostředí atd.). Algoritmická transparentnost je klíčová k budování důvěry veřejnosti k systémům UI. Technická řešení této problematiky se v současné době stále vyvíjejí (např. v projektu HE, AutoFair).

- **Otevřená zdrojová data, open source.** Měly by být kódy dostupné v open-source knihovnách? Nevýhodou může být zhoršení možnosti kontroly jejich férového a bezpečného využití v jiném kontextu. Podle jiných by být dostupné měly, ale ne na úkor duševního vlastnictví jejich majitelů. Transparentnost podle profesora Nigela Shadbolta z Jesus College Oxford (HC 945, 2023, 22. 2., s. 36) ještě nutně neznamená absolutní sdílení všech detailů a ve slyšení v Poslanecké sněmovně Britského parlamentu připomíná, že existuje široká škála informačních zdrojů, které popisují, jak lze zajistit „vysvětlitelnost" (explainability, XAI) že je potřeba tuto oblast ještě lépe prozkoumat. Vybudovat „ekosystém důvěryhodných datasetů" je pro něj prvořadé a za nejdůležitější považuje dodržovat principy transparentnosti a zodpovědnosti.
- **Práva vyplývající z duševního vlastnictví a copyright** – diskutované téma roku 2023 díky nástupu ChatGPT technologie. Využití, či přímo zneužití, intelektuálního vlastnictví umělců – od filmových tvůrců, malířů, grafiků až po hudebníky - firmami jako OpenAI při trénování jejich UI modelů (Barnes, Koblin & Sperling, 2023).

snahou překládat tyto termíny, které vědecká komunita používá v jejich anglickém originále (pozn. autorky).

- **Odpovědnost za škody** – složité právní téma. Na úrovni EU je „preferovaným řešením přísná odpovědnost založená na objektivním principu a zároveň systém povinného pojištění. […] Většina související legislativy je na jednotlivých členských státech (Kolaříková & Horák, 2020, s. 42-43). Je rozdíl např. pokud škoda (finanční, ekonomická) vznikla kvůli algoritmu na burze, nebo z provozu dronu, robota a případně autonomního vozidla a ve kterém státě. Obecně musí oběť/poškozená strana prokázat spojitost mezi škodou a provozovatelem, výrobcem daného zařízení či technologie, což může být velmi nákladné (EC DG JC, 2019, s. 20).
- **Zaměstnanost** – je zřejmé, že některá povolání, obory a profese budou využitím UI více postiženy než jiné (viz předposlední bod a problematika spojená se zneužíváním uměleckých děl, produkce levných replik, automatizace). U jiných profesí se bude jednat o tzv. **upskilling** – lidé budou vykonávat hodnotnější a zajímavější práci, např. programátoři budou softwaroví architekti a UI bude provádět kódování. Tedy rutinní úkony budou automatizovány UI systémy. Otázka je, jak dlouho tento stav potrvá, než výkonnější UI v budoucnosti nahradí i ty složitější a zodpovědnější profese.[36] V mnohých profesích je UI využívána pro tzv. **augmentaci**, tedy rozšíření schopností člověka, zefektivnění jeho výkonu, ale pozitivní vliv má i na zvýšení jeho bezpečnosti. To se týká např. vojáků na misích, členů složek IZS, pilotů, atd. Některé profese nahrazuje již nyní, např. řidiče kamionů v USA nebo zaměstnance v zemědělství.

  Gmyrek, Berg & Bescond (2023), provedli výzkum pro ILO – Mezinárodní organizaci práce (Working paper č. 96), jak ChatGPT technologie může ovlivnit či přímo ohrozit jednotlivé profese. Závěry jsou

[36] Šír, G. (2024, 21. 3.) při diskusi skupiny Responsible AI (RAI) na Katedře počítačů FEL ČVUT.

optimistické (s. 18). ChatGPT automatizuje procesy a dovednosti, které nevyžadují moc znalostí a dovedností (skills).

Jako při předchozích industriálních revolucích však vznikají a vzniknou díky UI některá zcela nová odvětví. Pracovní příležitosti tedy budou růst a bude záležet na schopnostech zaměstnanců rekvalifikovat se a zcela změnit profesi, kariéru. Zaměstnání tzv. bílých límečků – manažerů, akademických pracovníků, administrativy jsou profese nejvíce exponované UI (OECD, 2023; Green, 2023, s. 111).

- **Mezinárodní koordinace** – bude diskutováno dále v rámci kapitoly 3.
- **Existenciální rizika -** Vědecká komunita se polarizuje na tzv. futuristy, proroky, vizionáře nebo „doomsdayers“[37], kteří se zaměřují na rizika vzdálenější budoucnosti (předpovídají vznik tzv. superinteligence), které někdy připomínají scénáře katastrofických filmů[38], např. Kurzweil, Bostrom, Harari, Tegmark (Coeckelbergh, 2020, s. 13) a na ty „ukotvenější“ (pokud lze říci realističtější), kteří poukazují na negativní dopady, které algoritmy a neregulace UI způsobují již dnes. Pokud se nevyřeší a nezavedou se systematické kontrolní mechanismy včetně legislativy usměrňující vývoj v oblasti UI, budoucnost lidské civilizace, jak ji známe dnes, bude v důsledku rozložení základních stavebních kamenů společnosti nedemokratickými procesy a fake news, umocněnými algoritmy UI, velmi chaotická, ne-li nemožná (Brauner & Chan, 2023, August 10). Existenciálními riziky, mimo jiné spojenými právě s vývojem

[37] Pejorativní výraz označující ty, kteří neustále mluví o konci světa/civilizace způsobeném UI (novými technologiemi).

[38] Nikdo však netuší jak vzdálené a jestli bude daný pokrok technicky vůbec možný. Ti největší vizionáři hovoří o rizicích vzniku **technologické singularity**, kdy technologickému pokroku již člověk nebude vůbec rozumět a bude probíhat bez možnosti jeho zásahu. UI se bude sama sebezdokonalovat – ať už ve formě algoritmů/digitální nebo i zapojením biologického materiálu – např. kombinací s lidským mozkem či jiným ztělesněním (embodyment). Naopak Bostrom věří v tzv. **transhumanismus**, tedy technické/biologické vylepšování člověka tak, že vznikne nový typ lidí, kteří budou nesmrtelní – tzv. Homo Deus (Coeckelbergh, 2023, 31-33).
Všechny tyto možnosti představují existenciální rizika pro dnešní civilizaci, tak jak ji známe.

UI, se zabývá Jan Kulveit (viz podkapitola 1.1.5 Interdisciplinarita). Jak sám uvádí, zabývá se „výzkumem rizik, která lidstvo z nejrůznějších důvodů podceňuje, což jsou třeba rizika spojená s rozvojem technologií“ (Hubálková, 2022).

Publikace projektu Umělá inteligence a lidská práva: rizika, příležitosti a regulace (Šmuclerová M., Král, L., Drchal, J. et al., 2023c, od s. 17) nabízí celou řadu pozitivních příkladů využití UI v kontextu lidských práv, což je také občas potřeba vyzdvihnout. Pokud se UI vyvine a nasadí ve spolupráci s cílovými skupinami, s využitím kvalitních, vyvážených datasetů a transparentních algoritmů, je možné vymyslet zajímavá řešení, která zvýší bezpečnost, kvalitu života, přispějí ke zkvalitnění životního prostředí, biodiverzity atd.

1.4 UI jako politické, bezpečnostní a strategické téma

UI patří mezi strategické technologie a její vývoj, podpůrná infrastruktura, související legislativa, kontrola a financování patří mezi oblasti, kde je nutná mezinárodní, evropská i národní součinnost. Zároveň je ale mezi jednotlivými politickými celky předmětem konkurenčního boje a vývoj UI technologií a systémů má dopad do bezpečnostních a obranných struktur a strategií jednotlivých zemí a jejich závazků vyplývajících např. z členství v Severoatlantické alianci (NATO, 2020, 2021).

Definici těchto strategických cílů, vymezení hodnot, klíčových společenských témat, průmyslových odvětví a vědních oborů, kterých se tyto strategie týkají, které podporují a rozvíjejí, můžeme nalézt ve strategických dokumentech Evropské unie a národních vlád týkajících se směřování VaVaI – Evropského výzkumného prostoru (ERA, 2023), Evropské inovační politiky (EP, 2023) a podpory výzkumu a inovací v České republice (EC, 2021) na současné a následující časové období.

Rokem 2024 končí současné období, v rámci kterého si Evropská komise prostřednictvím Genálního ředitelství pro výzkum a inovace (EC, DGRI, 2023) vytyčila 7 „vyšších" politických cílů, jejichž formulace a prioritizace vychází z nadnárodních strategií[39], např. z cílů udržitelného rozvoje (SDGs) definovaných OSN. Dalším politickým tématem podporujícím soběstačnost EU je podpora vývoje kvantových technologií, v rámci tzv. programu Evopské digitální dekády, s cíli do roku 2030.[40]

Dalším klíčovým dokumentem je Strategická agenda pro výzkum a inovace (SRIA) Evropského Cloudu pro open science (EC, DGRI, 2022). Umělou inteligencí, robotikou a kybernetickou bezpečností a důvěrou se zabývá Generální ředitelství pro komunikační sítě, obsah a technologie (DG Connect), které spravuje Program Digitální Evropa. „Program slouží jako investiční doplněk ke Strategii pro jednotný digitální trh, která stanovuje pevný rámec v oblasti digitální ekonomiky a společnosti. Záměrem programu je také vytvořit napříč Evropskou unií síť Evropských center pro digitální inovace (EDIH) se specializací na vysoce výkonnou výpočetní techniku (HPC), kybernetickou bezpečnost a umělou inteligenci (DotaceEU).

V České repubice byla na období 2021 – 2027 stanovena Národní výzkumná a inovační strategie pro inteligentní specializaci České republiky (Národní RIS3 strategie), která rovněž zahrnuje strategie týkající se digitalizace a umělé inteligence (následuje citace ze s. 15, MPO, RIS3):

---

[39] UI spadá pod cíl č. 2. Evropa připravená na digitální věk – tam patří např. Akt o digitálních službách, Akt o digitálních trzích, Evropský akt o čipech a nová pravidla pro UI - AI Act, tedy právní rámec pro UI (EC DGRI, 2020).
Tyto strategické dokumenty vyhotovila všechna generální ředitelství E. komise na období 2020 – 2024. UI se tam objevuje i v rámci obdobného dokumentu DG Connect, DG Digit (pro informatiku) a samostatný dokument týkající se zakládání JRC (Společných výzkumných center). Jejich hlavní výhodou pro vědce má být nezávislost na národních/vládních a privátních zájmech.
[40] https://digital-strategy.ec.europa.eu/en/policies/europes-digital-decade

- **Digitální Česko**: Národní RIS3 strategie je úzce provázána s jejím pilířem „Digitální ekonomika a společnost.“ (MPO, RIS3, s. 15)
- **Národní strategie umělé inteligence ČR 2019-2035** (MPO, 2019) schválená v květnu 2019, která představuje základní strategický dokument pro rozvoj UI v ČR. Strategie pokrývá širokou škálu oblastí od podpory vědy a výzkumu přes vzdělávání po problematiku regulace a mezinárodní spolupráce v UI.
- **Strategický rámec Svazu měst a obcí v oblasti Smart City:** zaměřený na rozvoj ČR v oblasti Smart City/Smart Region (SMOCR, 2020).

V současné době probíhá revize Národní strategie pro UI na MPO. V únoru 2024 proběhl mezinárodní kulatý stůl s experty z Velké Británie, USA, Nizozemska a Kanady (MPO, 2024):

„Debata se zaměřovala zejména na klíčové iniciativy a pilíře, zapojení hospodářských a sociálních partnerů do implementace strategií, lákání talentů či vládní přístup k rozvoji AI. Všechny účastnické země workshopu potvrdily důležitost prosazování umělé inteligence se zaměřením na člověka (AI alignment), prosazování přístupu s ohledem na možná rizika, která tato technologie představuje. Účastníci také zdůraznili nutnost podpory mezinárodní spolupráce v oblasti umělé inteligence.”

V rámci aktualizace Národní strategie vyplynulo z veřejné konzultace s 517[41] respondenty (MPO, 2023, říjen) mimo jiné i to, že: ...

[41] Téměř polovinu respondentů tvořili zástupci veřejnosti, 19 % zástupci státní správy, 16 % ze soukromé, 13 % akademické sféry.

„… etické využívání a právní rámec pro umělou inteligenci a posilování základních dovedností a vzdělávání v oblasti umělé inteligence by podle respondentů ***měly patřit mezi klíčové oblasti aktualizované Národní strategie umělé inteligence.“***

Etika využívání UI byla na 3. místě s 201 hlasy a právní rámec pro UI na 4. místě se 178 hlasy.[42] Nejvíce respodentů hlasovalo pro finanční podporu kyberbezpečnosti (239 hlasy).

Nutnost zodpovědného a strategického přístupu k novým technologiím a zejména k UI systémům vyplývá i z Výroční zprávy 2022 Bezpečnostní a informační služby (BIS, 2023). Mezinárodním bezpečnostním tématem jsou autonomní zbraně. V případě autonomizace zbraní prostřednictvím využití UI dochází k tzv. digitální dehumanizaci.[43] Válka probíhá na dálku, bez přímé fyzické účasti útočníka a jejího plného efektu na obě strany konfliktu, což s sebou nese mnoho morálních dilemat. O nich pojednává např. dokument „The Heart of Immoral Code“.[44]

NATO oznámilo v srpnu 2023 založení nového inovačního fondu (NIF) o velikosti 1 mld. EUR. Česká republika patří mezi 23 spojeneckých států, které fond založily a budou jej moct využít k podpoře místní start-upové scény v oblasti výzkumu vojenských technologií. Dalším fondem, kterým NATO podpoří vznik univerzitních spin-offů, je DIANA Defence Innovation

---

[42] Respondenti mohli vybírat až pět preferovaných odpovědí.

[43] Známá je Kampaň k zastavení robotů zabijáků – The Campaign to Stop Killer Robots. https://www.stopkillerrobots.org/news/unban/ Jejím cílem je tlačit na politickou reprezentaci, aby zabránila automatizaci a externalizaci zabíjení. Podobné iniciativy jsou a v minulosti byly i v oblasti vývoje biologických a chemických, a po druhé světové válce i nukleárních zbraní.

[44] Ukázka reakcí respondentů na některá témata a otázky, týkající se morálních zásad a rozhodnutí, jsou k dispozici zde: https://immoralcode.io/index.html. Ačkoliv se jedná o základní otázky, odpovědi již nejsou tak jednoznačné, např. na otázku: Byli byste schopní zabít člověka? Ačkoliv někteří respondenti rázně odpovídají, že ne, mnozí váhají – záleží na okolnostech, v případě, že by je někdo napadl, atd. „A kdyby bylo útočníkem 12leté dítě?“
V případě autonomních zbraní se mentálně dostáváme spíš do oblasti simulací a gamingu. Jednotlivé cíle jsou dehumanizovány, odlidštěny. Nejsme s nimi v přímém kontaktu, neohrožují bezprostředně život konkrétního útočníka. Další oblastí rizika je možnost zneužití těchto autonomních zbraní, útoku hackerů, ale i jejich budoucí využití protistranou. Měla by se otázka života a smrti rovnat jedničkám a nulám algoritmu, který o tom rozhodne za nás?, táže se zmíněný dokument.

Accelerator for the North Atlantic. Obory, které DIANA podpoří, budou UI, autonomie, kvantové a bio-technologie, lidské vylepšování (enhancement), hypersonické systémy, vesmír, nové materiály a výroba, energie a pohon a komunikační sítě nové generace (Vodička, MO CR, 2023).

Další oblastí, která představuje mezinárodní bezpečnostní rizika a je spojená s UI a využitím velkých dat/ big data, je oblast tzv. pozorování Země. Za pomoci čím dál dokonalejších družic, které, kromě např. ESA mnohdy patří korporacím, a zobrazovací techniky s vysokým rozlišením a schopností ukládání velkého množství dat a jejich algoritmického vyhodnocování, je možné dlouhodobě sledovat i strategické cíle, přesuny a logistiku v oblasti strategických zbraní, vojenské techniky a surovin, budování vojenských základen a logistických center, které jsou významné pro národní obranu (AI4EO, 2024).

Pozitivním využitím těchto technologií a získaných dat je potom např. v rámci Platformy Google Maps. Ta od září 2023 poskytuje firmám a vývojářům přes rozhraní API[45] možnost využívat environmentální data a letecké snímky, které mapují vhodné lokality pro využití solární energie (vychází to z projektu Google – Sunroof), pylové znečištění a kvalitu ovzduší (Maguire, 2023), aby sami mohli vytvářet aplikace v oblasti životního prostředí.

Pokrok ve vývoji UI a jeho dopad na společnost, ekonomiku, průmysl, vzdělávání, životní prostředí i veřejné mínění popisuje každý rok zpráva „AI Index“. Jedná se o nezávislou iniciativu „Stanfordského institutu pro UI zaměřenou na člověka“. Tato zpráva nejen sleduje a shrnuje nejnovější vývoj v oblasti UI, který je striktně založen na datové analýze, ale zároveň ovlivňuje i budoucí vývoj a směřování v oblasti UI. Její výsledky totiž sledují investoři, kteří

[45] API (application programming interface – aplikační programovací rozhraní) je: „Soubor postupů, funkcí, knihoven, protokolů a tříd, které umožňují dvěma aplikacím vyměňovat si mezi sebou data, a tedy komunikovat jednosměrně či obousměrně. API rozhraní zpřístupňuje programátorům informace z jiných aplikací, které jsou poté využity k rozšíření funkcionality webu nebo automatizaci některých procedur. Příkladem tak může být například zobrazení polohy na Google mapách vložené na webových stránkách nebo automatické vygenerování faktury.“ https://idealab.cz/slovnik/api-rozhrani/

se zajímají o oblast nových technologií a technologických startupů, ale i akademická sféra a státní správa.

Očekávání, investice a schopnost vědců/vědkyň v akademické či privátní sféře dodat předpokládané výsledky ovlivňuje rychlost vývoje a směřování oboru UI od jeho počátku. Růstové spurty spojené s přelomovými pokroky ve vývoji UI znamenají vyšší očekávání i vyšší investice. Pokud další fáze vývoje trvá neúměrně dlouho a investoři jsou zklamáni, nastává zpomalení, až zadrhnutí vývoje, mluvíme potom o tzv. „UI zimě".

2 Výzkumný ekosystém v oblasti výzkumu, vývoje a inovací umělé inteligence (VaVaI UI)

Ekosystém VaVaI pro UI v České republice je, zjednodušeně, prostor, který je legislativně vymezen (EU a ČR legislativou, mezinárodními smlouvami a úmluvami), probíhá v něm nabídka a poptávka po výzkumu **ve veřejném a strategickém zájmu ČR (a EU),** viz předchozí podkapitola, ale i vývoj inovací, který je podporován či vykonáván veřejnými výzkumnými institucemi a soukromými subjekty. Působí zde **aktéři**, kteří spolupracují, ale zároveň si i konkurují (výzkumné organizace – univerzity, Akademie věd ČR, soukromé subjekty).

Partnerství, spolupráce, kooperace mezi jednotlivými subjekty na různých úrovních moci a funkce v rámci tohoto „ekosystému" jsou nezbytné k dosažení stanovených cílů (strategických, politických, ekonomických, společenských a dalších, např. cílů v oblasti zdravotnictví, životního prostředí a zemědělství). Na nejvyšší, globální úrovni, to jsou potom „společné cíle, které byly přijaty na Summitu OSN v září 2015 a jsou společným rozvojovým programem všech států světa, cíle udržitelného rozvoje (SDGs) (OSN – UN, 2024)."

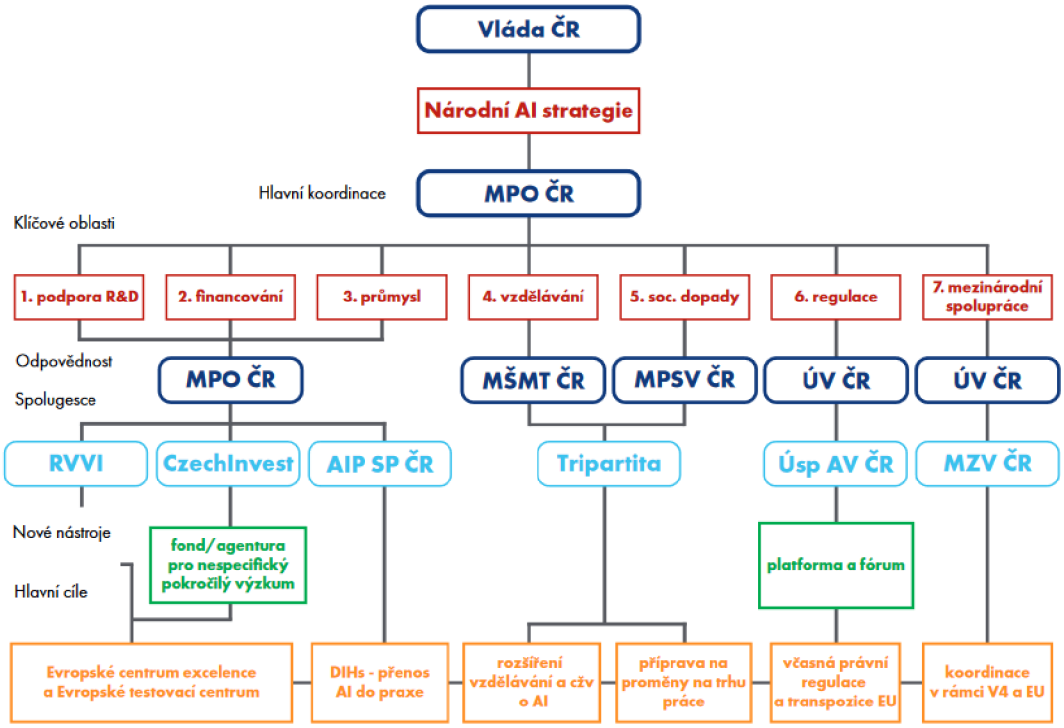
*„**Inovační strategie** České republiky 2019–2030 byla schválena Usnesením vlády ČR ze dne 4. února 2019 č. 104. Jedná se o strategický rámcový plán, který předurčuje vládní politiku v oblasti výzkumu, vývoje a inovací […]. Inovační strategie se skládá z devíti navzájem provázaných pilířů, které obsahují východiska, základní strategické cíle a nástroje vedoucí k jejich naplnění. Jsou jimi oblasti: **Financování a hodnocení výzkumu a vývoje**, **Inovační a výzkumná centra**, Národní start-up a spin-off prostředí, **Polytechnické vzdělávání**, Digitalizace, Mobilita a stavební prostředí, Ochrana duševního vlastnictví, Chytré investice a Chytrý marketing"* (RVVI, 2019).

Rozvoj UI, digitalizace a nových technologií je některými z těchto pilířů přímo podporován, jinde mohou být UI technologie využity jako nástroje k naplňování strategických cílů těchto pilířů. Vláda ČR deklarovala, že podpoří investice „v oblastech klíčových trendů, dle Strategie chytré specializace (RIS3), Národního kosmického plánu a Strategie podpory umělé inteligence."

Tyto strategické dokumenty vymezují směřování České republiky nejen v oblasti výzkumných témat a priorit, ale specifikují i praktické **nástroje a mechanismy**, jak naplnění těchto cílů podpořit a zajistit. To ovlivňuje pravidla ekosystému VaVaI UI na celé toto období.

Důležitou podmínkou a cílem by měl být i **důraz na kvalitativní stránku technických řešení využívajících UI** s ohledem na bezpečnost, transparentnost, vysvětlitelnost a další principy odpovědné UI a vzít v úvahu jejich možný společenský vliv/dopad. Dokument Národní AI strategie (NAIS) vyzdvihuje důležitost zaměřit excelentní výzkum na vývoj „odpovědného a důvěryhodného AI (NAIS, 2019, s. 8)." UI má být, podle dokumentu, také vysvětlitelná. Strategie si klade za cíl i: „edukaci a celospolečenskou osvětu v oblasti regulace a etiky AI, včetně podpory interdisciplinárních studijních oborů (s. 35)."

Další koncepcí, jež spadá do gesce MPO, které bude v roli hlavního koordinátora – viz obrázek č. 6 - je „Digitální ekonomika a společnost“ (Dzurilla, Očko et al., 2020). Hlavním cílem tohoto dokumentu je: „nastavit funkční a flexibilní právní, finanční a institucionální rámec tak, aby posílil konkurenceschopnost a zároveň pomohl předejít negativním dopadům digitální transformace na společnost.“



Obrázek 6 Diagram vazeb subjektů, cílů a nástrojů implementace Národní AI strategie v ČR (NAIS, 2019, s.14).

Deklarovanými cíli jsou (Dzurilla, Očko et al., 2020, s. 3):

1. Efektivnější systém přímé i nepřímé podpory VaVaI
2. Zralost a připravenost sektorů na digitální transformaci
3. Připravenost občanů na změny trhu práce, vzdělávání a rozvoj digitálních dovedností
4. Podpora konektivity a infrastruktury digitální ekonomiky a společnosti
5. Zajištění bezpečnosti a důvěry v prostředí digitální ekonomiky a společnosti

6. Legislativa podporující všechny aspekty digitální ekonomiky a společnosti
7. Optimální systém financování digitální ekonomiky a společnosti
8. Institucionální zajištění centrální koordinace politik na podporu digitální ekonomiky a společnosti

V novější verzi vypracované I. Bartošem et al. (2023) je změna v bodě 1. Je tam nově text: „Podpora výzkumu, vývoje a inovací v oblasti digitální ekonomiky a společnosti“ a je vynechán bod č. 7.[46]

Pokud se tedy vrátím k výzkumné otázce: Jaké kontrolní mechanismy ovlivňují výzkum a výstupy v projektech využívajících nebo vyvíjejících umělou inteligenci (AI), aby byly v souladu s etickými principy?, přemýšlím, zda cíle těchto dokumentů ve vztahu k výzkumu a vývoji UI nejsou příliš vágní a jak jsou deklarované cíle naplňovány.

2.1 Mechanismy

Mechanismy, kterými vláda ČR ovlivňuje ekosystém VaVaI UI a přímo je v dokumentech zmiňuje jsou: financování a reforma hodnocení vědy a výzkumu a zaměření se na podporu polytechnického vzdělávání. Dalším důležitým mechanismem je podpora inovací, zapojování studentů do výzkumu a zakládání Inovačních a výzkumných center.

Mechanismy jsou jednotlivé agendy, procesy, které umožňují dosáhnout cílených změn působením/ovlivňováním chování aktérů v rámci ekosystému VaVaI UI. Toto ovlivňování je motivačně-regulační ve smyslu odměn (finance, lepší podmínky pro studium, výzkum, mobilitu, kladné hodnocení při splnění určitých podmínek atd.) a omezení (nastavení pravidel čerpání výzkumných grantů právě tak, aby bylo dosaženo změny. Např. že bude ve výzkumu kladen

[46] V dokumentu jsou další změny, vzhledem k uzávěrce už ale nebyly do DP zapracovány.

větší a jasně prokazatelný důraz na etické aspekty UI ve všech fázích životního cyklu, že budou podporovány určité – transparentní - výzkumné metody, etické sebehodnocení bude vyžadováno jako součást grantové přihlášky atd.).

Vláda a ministerstva tedy aktivně VaVaI na jednu stranu umožňují a podporují prostřednictvím grantové politiky a vytvářením prognóz a strategií v oblasti nových technologií a stanovením dalších ekonomicko-sociálních cílů, ale zároveň fungují v roli regulátora. V rámci jednotlivých mechanismů tedy určí už konkrétnější nástroje. Příkladem může být nastavení konkrétních pravidel čerpání a vykazování nákladů (včetně toho, jak bude administrativně náročné vykazování nákladů v realizační fázi výzkumného projektu. Pokud potom srovnáme režim lump sum u HE a povinnost denních výkazů práce v českých podmínkách nebo % výše povinného kofinancování atd., jsou to také určité signály). Dále má určité podpůrné nástroje s cílem zajistit, aby se vědcům dobře pracovalo, aby mohli vybudovat dobrý tým a přilákat špičkové vědce nebo nadějné doktorandy a postdoktorandy ze zahraničí atd.

Níže následuje přehled jednotlivých aktérů, mechanismů, podmínek a pravidel ekosystému VaVaI UI. Seznam jistě není zcela kompletní. Prezentuji ho zde jako součást mého empirického výzkumu založeného na rozhovorech s vědci a managementem a vlastní praxe manažerky výzkumných projektů.

Uvedený seznam aktérů a mechanismů dokazují, jak komplexní prostředí ekosystém VaVaI UI je. V mnohém je to prostředí stejné pro všechny obory, ale podmínky, příležitosti a omezení vyplývají z výše uvedenýc strategických plánů a cílů. Stejně tak s každou specializací jsou spojena jiná rizika, zákony a povinnosti i příležitosti a podle toho se potom liší i složení dalších aktérů v tomto ekosystému.

2.1.1 Základní členění mechanismů

Mechanismy bych rozdělila na **formální** (dané zákonem, hierarchií podle výkonné a legislativní moci) **a neformální** (vycházejí z kultury, praxe, tradic

určitého oboru nebo vědecké komunity, práce dobrého vědeckého vedoucího výzkumné skupiny může pozitivně ovlivnit několik generací mladých vědců a vědkyň), **uplatňované shora** (top-down) nebo jako **tlak zdola** (bottom-up), jako **externí motivátor** (povinnost, regulace, podpora, incentiva) nebo **interní motivátor** (profesní vědecká integrita a morálka – RRI – viz kapitola 3, touha dělat excelentní vědu bez zkratek a s ohledem na ostatní, které chování vědce a výsledky jeho výzkumu přímo či nepřímo ovlivňuje. Tedy uvědomění si širších souvislostí vlastní práce).

Patří sem i **faktor času** a další vlivy, které lze těžko predikovat. Některé plánované změny a výsledky politik se projeví za delší dobu než se očekávalo, jiné dříve, technologie se vyvinou jiným směrem než se plánovalo (viz ChatGPT v roce 2023, který překvapil i mnohé vědce z oblasti UI), neznámý je i budoucí vývoj geopolitické situace ve světě. V nedávné době proběhla i celosvětová pandemie Covid-19, která významně ovlivnila celý ekosystém VaVaI.

Všechny tyto vlivy na sebe vzájemně působí. Je nutné neztratit orientaci, co je v daný moment důležité. To musí koordinovat globální a mezinárodní instituce a národní vlády, které by měly vypracovat i nový plán managementu rizik a, jak uvádí Korinek (in Dubber, Pasquale & Das, 2020, s. 475-6), nepodléhat techno-fatalismu ani přehnanému techno-optimismu.

2.1.2 Aktéři

"Software" UI VaVaI – tedy znalostní a výzkumná centra a lidský kapitál: EU podporuje kooperaci různých aktérů na strategických tématech – např. formou vzniku různých klastrů a center excelence: European Quantum Excellence Centres (QECs) v oblasti kvantové informatiky, nebo viz. kapitola 1, GenAI4EU atd. UI je zmiňována prakticky ve všech strategických dokumentech

– v oblasti rozvoje, ale i nutnosti přijmout opatření proti negativním dopadům UI na společnost.[47]

Výzkumné veřejné nebo státní instituce[48] mají ze Zákona o VŠ č. 111/1998 Sb. i další funkci v oblasti etiky, podle paragrafu 1, zejména v odstavci d). „Vysoké školy jako nejvyšší článek vzdělávací soustavy jsou vrcholnými centry vzdělanosti, nezávislého poznání a tvůrčí činnosti a mají klíčovou úlohu ve vědeckém, kulturním, sociálním a ekonomickém rozvoji společnosti tím, že: d) hrají aktivní roli ve veřejné diskusi o společenských a etických otázkách, při pěstování kulturní rozmanitosti a vzájemného porozumění, při utváření občanské společnosti a přípravě mladých lidí pro život v ní.“ Každá univerzita a vysoká škola realizuje tuto svou úlohu jiným způsobem a do jiné míry a intenzity, se zapojením různých cílových skupin a různou mírou a oblastmi společenského dopadu.

2.1.3 **Strategie, Legislativa, Kontrola/Audit, Hodnocení vědy:**

Vláda a ministerstva (Ministryně pro vědu, výzkum a inovace; MŠMT; Rada pro výzkum, vývoj a inovace RVVI při Úřadu vlády ČR, MD, MO, MZ, GA ČR, TA ČR atd.) „Legislativa" (jako soft law) i ve formě specifikace podmínek a pravidel pro získání vědeckých grantů. Systém hodnocení vědy je velké téma, které ovlivňuje výzkumnou kariéru a strategii vědců. Kvantitativní přístup byl, minimálně na úrovni, Evropské rady pro výzkum (ERC) překonán a od letošního roku bude hodnotit výsledky v souladu s Dohodou o reformě hodnocení vědy (CoARA, 2022). Této dohodě předcházely iniciativy jako DORA (Declaration on Research Assessment), Hongkongské principy a další.

---

[47] Projekt Future-pro: megatrendy a velké společenské výzvy identifikoval Etiku UI jako výzvu č. 16 v rámci kapitoly Trendy a výzvy v obasti: 10 Hodnoty. Dalšími riziky souvisejícími s UI byla kyberbezpečnost a a "automatizace lidské práce a jejich negativní vliv na různé segment ekonomiky, prohloubení ekonomické nerovnosti mezi státy a rostoucí technologická nezaměstnanost." https://www.megatrendy.cz/

[48] Podle Zákona o vysokých školách a o změně a doplnění dalších zákonů, č. 111/1998 Sb.,je státní vysoká škola vojenská nebo policejní. Zřizuje je přímo Ministerstvo obrany ČR a Ministerstvo vnitra ČR.

2.1.4 **Finanční toky**

(mezi aktéry VaVaI) jsou multizdrojové: kombinuje se zde podpora z EU, ČR rozpočtu, soukromých investic, ale i jiné zahraniční investice a granty. Tyto finance jsou administrovány přímo Evropskou komisí, Ministerstvy nebo přes agentury GA ČR, TA ČR, atd.

2.1.5 **Výzkum, vývoj, inovace, vzdělávání, podpora**

Zajišťují je výzkumné organizace, AV ČR, univerzity, Centra transferu technologií, Úřad průmyslového vlastnictví, Technologické centrum Praha, ale i neziskové organizace jako prg.ai, které podporují místní výzkumný ekosystém v oblasti UI a propojují výzkumné organizace mezi sebou, např. vytvářením společných bakalářských a magisterských studijních programů, napojením na Pražský inovační institut a další aktéry z řad samosprávy a státní správy, startupovou scénu a korporace.

2.1.6 Rada Evropy

připravuje vlastní „(Rámcovou) Úmluvu o UI“ (Artificial Intelligence, Human Rights, Democracy and the Rule of Law Framework Convention). Konečná podoba byla zfinalizována v březnu 2024 a bude postoupena Výboru ministrů při Radě Evropy k podpisu v následujících týdnech. Jedná se o dokument, který nastavuje vysoké etické standardy a budou se k němu moci připojit země z celého světa.

2.1.7 Evropská Unie

Evropská komise dne 21. 4. 2021 vydala balíček iniciativ (tzv. AI balíček), který obsahuje aktualizovaný Koordinovaný plán k AI, který se věnuje podpoře talentů a digitálních dovedností, investicím do výzkumu, vývoje a využití AI v klíčových oblastech (např. zemědělství, životní prostředí, mobilita, veřejná správa nebo zdravotnictví), a dále obsahuje návrh nařízení, kterým se stanoví harmonizovaná pravidla pro UI na vnitřním trhu (tzv. AI Act). Téma umělé inteligence patří také mezi prioritní témata Víceletého finančního rámce pro

období 2021–2027 (např. jde o programy Horizont Evropa, Digitální Evropa, Národní plán obnovy) (MPO, 2024).

Ze závěrů konference projektu UI a lidská práva: Rizika a příležitosti (září 2023) vyplynulo, že kdyby byla legislativa GDPR správně implementována do národních legislativ, z právního pohledu by to ošetřilo většinu případů spojených s negativním dopadem UI na oblast lidských práv (Šmuclerová M., Král, L., Drchal, J. et al. , 2023, 21. 9.). To klade velké nároky i na způsob, jak bude nakonec implementován i AI Akt v Českém prostředí, aby byla zajištěna jeho efektivita. Diskuse byla i na téma, zda bude zřízen samostatný úřad pro jeho zavedení, správu a kontrolu – na úrovni EU byla v únoru 2024 otevřena Kancelář UI (AI Office), a jak dojde k přerozdělení agendy spojené s AI Akt mezi případný nový úřad/orgán a stávající úřady.

Koordinaci zajistí Odbor koordinace digitální agendy, pod Odborem věcných politik EU, zajišťuje výkon činností spojených s koordinací politiky jednotného digitálního trhu v Evropě (Vláda, 2024), pravděpodobně spolu s Digitální a informační agenturou (DIA, 2024) a Národním úřadem pro kybernetickou a informační bezpečenost (NÚKIB, 2024). Podobu implementace AI Akt v jednotlivých členských státech EU projednává Rada EU. AI Akt se bude týkat i spotřebitelské legislativy, kterou má v gesci MPO, kybernetické bezpečnosti, ta je v gesci MV ČR (např. Akt o kybernetické odolnosti).

Na úrovni EU má mandát k zajištění regulačního rámce kybernetické bezpečnosti agentura ENISA. Nejcitlivějšími sektory, kde bude legislativa AI Akt implementována budou letectví, doprava, lékařství a finanční sektor. AI Akt v tomto případě doplní ještě sektorová legislativa (Šmuclerová M., Král, L., Drchal, J. et al., 2023, 21. 9.). UI systémy musí být posuzovány z hlediska celého životního cyklu. Je nezbytné stanovit potenciální oblasti ohrožení (lidských) práv hned na začátku (viz také Ethics-by-design přístup – v podkapitole 3.4). AI Akt podle autorů dostatečně nerozlišuje rozdíl mezi transparentností a vysvětlitelností – viz podkapitola 1.3

AI Akt bude tzv. performativní regulace. UI systémy jsou rozřazeny do kategorií podle jejich vlivu na společnost – tedy míry rizika. Ty, které spadají do kategorie „Nepřijatelné riziko“, viz níž, jsou přímo zakázané. Kategorie „vysoké riziko“ bude řízena zvláštními podmínkami (Šmuclerová M., Král, L., Drchal, J. et al. , 2023, 21. 9.). Tento systém bude muset projít posuzováním shody a splnit příslušné požadavky. Bude zaregistrován v databázi EU a bude muset být podepsáno prohlášení o shodě, systém bude opatřen označením CE. Teprve potom bude možné tento systém uvést na trh a v případě změny jakékoliv z podmínek, za které byl ocertifikován, bude muset projít celým procesem znovu. Uživatelé budou hlásit závažné incidenty a malfunkce (EC, 2024).[49]

AI Act bude podle Jakuba Marečka (2024):

„omezovat algoritmy, jejichž chování může být viděno jako adaptivní. Tzn., že na základě dat nebo vstupů, které dostávají, mohou v průběhu času dávat různé výsledky. „V tom je ta regulace velmi široká a definice toho, co je UI je takto asi nejšířeji pojatá z definic, které se uvažovaly v průběhu vyjednávání o Aktu o UI.“

Rozdělení systémů podle rizik je podle názoru J. Marečka (2024) nové hlavně v tom, že zavádí novou regulaci pro generativní UI a pro systemicky důležité systémy generativní UI.

„Vymezení těch systemicky důležitých systémů je obsaženo v návrhu AI Aktu. Tam je celá řada podmínek, ať už v počtu operací, které byly použity pro učení toho systému, nebo v počtu firem, které je používají v EU atd.“

Podle názoru Marečka (2024) jsou: ...

---

[49]Základní dokumenty EU k UI (viz Evropská komise, KOORDINOVANÝ PLÁN V OBLASTI UMĚLÉ INTELIGENCE – PŘEZKUM 2021):
„Evropská komise, sdělení Koordinovaný plán v oblasti umělé inteligence (COM(2018) 795 final). Evropský přístup k UI, včetně hodnot, na jejichž základě chce EU pokročit v rozvoji a zavádění UI, je stanoven ve sdělení Evropské komise Umělá inteligence pro Evropu (COM(2018) 237 final) a v dokumentu Bílá kniha o umělé inteligenci – evropský přístup k excelenci a důvěře (COM(2020) 65 final).
Declaration of Cooperation on Artificial Intelligence (Prohlášení o spolupráci v oblasti umělé inteligence), podepsané všemi členskými státy a Norskem, duben 2018 (EC, 2024).“

„... ty systemicky důležité systémy (generativní UI) vymezeny poměrně neprůstřelně. […] A pak je tam nová kategorie vysoce rizikových systémů a tam asi uvidíme diskuse a judikáty.“

Nepřijatelné riziko – zakázané praktiky (Státní úřad inspekce práce, 2023):

„**Manipulace:** Systémy umělé inteligence, které manipulují s chováním lidí nebo specifických zranitelných skupin (např. hračky aktivované hlasem, které podněcují nebezpečné chování u dětí). **Sociální skórování:** Klasifikace lidí na základě chování, socioekonomického statusu nebo osobních charakteristik. **Biometrická identifikace a kategorizace:** Včetně reálného a vzdáleného biometrického identifikačního systému, jako je rozpoznávání obličeje v reálném čase (Státní úřad inspekce práce, 2023).“

Na úrovni EU bude ustanovena Evropská rada pro UI („Rada“):[50] ve které bude zasedat vždy jeden zástupce za každý členský stát. Jako pozorovatel se jednání Rady bude účastnit Evropský inspektor ochrany údajů a Kancelář UI (The AI Office) bez hlasovacího práva, případně další národní a evropské autority podle konkrétní agendy a potřeby, pokud to bude relevantní (FLI, 2024).

„Platnost a aplikovatelnost (Státní úřad inspekce práce, 2023):

AI Akt bude formálně přijat v dubnu 2024 a bude plně aplikovatelný 24 měsíců po vstupu v platnost. Zakázané systémy umělé inteligence musí být vyřazeny 6 měsíců po vstupu aktu v platnost. Kódy praxe (soubory dobrovolných pokynů, standardů a nejlepších postupů) se stanou aplikovatelnými 9 měsíců po vstupu v platnost.

Pravidla pro systémy obecného účelu, které musí splňovat požadavky na transparentnost, se stanou aplikovatelnými 12 měsíců po vstupu v platnost. Povinnosti týkající se vysoce rizikových systémů se stanou aplikovatelnými 36 měsíců po vstupu aktu v platnost.

[50] A ‘European Artificial Intelligence Board’ (the ‘Board’)

2.1.8 Publikační domy, databáze WoS a SCOPUS, redakce vědeckých žurnálů

ovlivňují kulturu nastavením publikačních podmínek, jestli vyžadují etické sebehodnocení, atd.

2.1.9 Etické výzkumné komise

Práce etických komisí ve vědeckých institucích TOP-down (v případě přístupu Vídeňské univerzity i bottom-up – příklad v kapitole 3)

2.1.10 Členství v mezinárodních organizacích

Zavázání se k dodržování etických pravidel vyplývajících z **členství** v mezinárodních i místních organizacích sdružujících výzkumné organizace, technické a jiné univerzity atd. – příklady **SEFI, EUA.** Většinou má formu etického kodexu.

2.1.11 Oborové organizace

Pravidla etiky UI také definují jednotlivé **oborové organizace** v oblasti výpočetní techniky, informačních technologií, umělé inteligence, strojového učení a dalších oborů. Největší profesní organizace sdružující výzkumníky a odborníky těchto inženýrských oborů jsou: **IEEE, AAAI , ACM** (Asociace pro výpočetní techniku), které jsou zodpovědné za definici standardů a pravidel pro jednotlivé výzkumné obory, organizaci mezinárodních vědeckých oborových konferencí a zajištění kvality publikovaných vědeckých článků v rámci jejich vlastních vědeckých žurnálů či konferenčních sborníků. O rostoucím významu a možnostech pozitivně ovlivnit další vývoj v oblasti etiky UI změnou podmínek pro publikování a přijetí příspěvků na výzkumné konference, pojednává článek Srikumar et al. (2022, s. 1061-1064).

IEEE vydala rozsáhlou publikaci **Ethically Aligned Design**, jedná se o skutečně globální iniciativu s přispěvateli a komisemi na různá odborná témata

složené z přispěvatelů z celého světa. Poslední verze byla publikována v roce 2019 na základě reeditace dvou předchozích s pomocí odborné veřejnosti. Etický design vztahuje IEEE k pěti obecným principům: Lidská práva, Celková pohoda (well-being), Odpovědnost (accountability), Transparentnost, Uvědomělost. V dokumentu je dále podrobně rozvedeno, jakými kroky těchto cílů dosáhnout a naplnit je. AAAI (Asociace pro pokrok umělé inteligence – Association for the Advancement of AI) má také vlastní Codex profesní etiky a jednání (conduct).

2.1.12 Standardizační organizace –

CEN-CENELEK, International Organization for Standardization and the International Electrotechnical Commission (**ISO/IEC), ETSI či NIST** ve Spojených státech Vypracovávají mezinárodní standardy včetně specializace a praxe UI.

Ty mohou být definovány čistě technicky a kodifikovány do podoby normy/standardizace daného oboru. Standardizační agenda v oblasti UI je velmi rozsáhlá. Pro každou z technologií založených na UI či využívajících UI byla v rámci ISO zřízena vlastní pracovní skupina – např. pro technologii zpracování přirozeného jazyka, funkční bezpečnost sytémů UI, pro oblast principu důvěryhodnonosti (trustworhiness) a mnoho dalších.

Významným standardizačním úřadem pro standardizaci umělé inteligence je VDE a DKE, Německé komise pro elektrické, elektronické a informační technologie v rámci Německého institutu pro standardizaci (DIN).[51]

2.1.13 Organizace a management výzkumných projektů

Agenda RMA (Research management and administration) – pravidla pro podávání projektových žádostí ve fázi pre-award, tedy přípravy projektu po vědecké a technicko-administrativní ovlivňují podobu konečné žádosti a

[51] Tyto informace jsem získala a publikuji s informovaným souhlasem JM. Rozhovory proběhly v červenci 2023. Informace jsem ověřila na webových stránkách přílušných organizací.

projektu a následně i jeho realizaci (post-award). Viz etické sebehodnocení, příručky, metodika Ethics-by-Design v podkapitole 3.4.

2.1.14 Neformální mechanismy

Procesy a chování ovlivněné neformálními faktory: jako tradice, kultura a chování v určitém oboru, instituci, kontextu. Vliv tradiční hierarchické kultury vědeckých institucí atd.

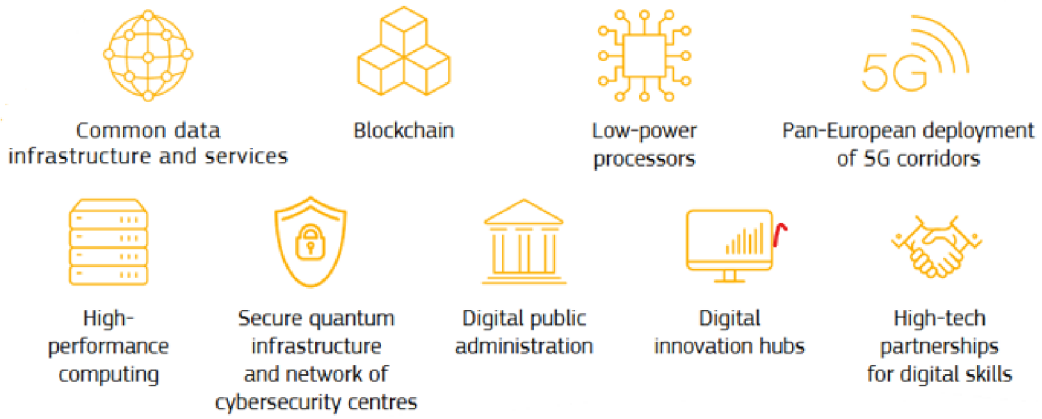
Vliv má např. i stále převážně maskulinní kultura oborů spojených s vývojem a výzkumem v oblasti UI. Toto téma se prolíná celou DP - např. v podkapitole 1.1.1. Inteligence nebo již zmíněný termín „technošovinismus" v podkapitole 1.2. Tato maskulinní kultura pravděpodobně ovlivňuje celkový přístup oboru UI k tzv. měkkým tématům, za jaké je etika UI považována. Neznamená to však, že by v oblasti UI nepůsobily výrazné osobnosti vědkyň jako např. Joanna Bryson, Virginia Dignum, Mia Shah-Dand, Joy Buolamwini či Timnit Gebru, a pochopitelně i na českých technických vysokých školách. Mnohé z nich se sdružují ve WAIE – The Women in AI Ethics a svou činností zároveň zviditelňují i téma etiky UI, genderové rovnosti nebo „zodpovědné UI“ a související aktuální témata.

2.1.15 Dostupnost vybavení, kvalita infrastruktury

„**Hardware" UI VaVaI**, tedy nezbytná infrastruktura potřebná k výzkumu a digitalizaci správy a řízení, je ve vlastnictví státu, soukromých subjektů nebo výzkumných organizací. Jedná se o různé telekomunikační páteřní sítě (zajišťované např. CETIN), kvantové počítače (viz konsorcium LUMI-Q, který bude umístěn na TU Ostrava a integrován do EuroHPC superpočítače KAROLINA)[52] výpočetní klastry (HPC – High-Performance Computing), laboratoře, výzkumnou infrastrukturu, ale i centralizaci strategické výroby – polovodiče, čipy, procesory na území EU (EuroHPC, 2024).

[52] https://eurohpc-ju.europa.eu/eurohpc-ju-launches-procurement-lumi-q-2024-02-14_en

Ve strategickém dokumentu EU komise – Digitální dekáda do roku 2030 je tato nutná infrastruktura blíže popsána, převážně v oblasti zájmů, do kterých technologií a podpůrné infrastruktury a systému plánuje EU investovat– viz. obrázek č. 7. Český ekosystém VaVaI UI nemůže fungovat nezávisle na partnerech z EU. Specifické oblasti, které EU definovala jako kooperativní (multi-country) a finančně je podpoří jsou (EC, 2023):



Obrázek 7 Plánované strategické investice do výzkumné infrastruktury a dalších strategických oblastí v rámci digitalizace veřejné správy (EC, 2023, s.3)

Všichni aktéři interagují v závislosti na daném procesu, typu výzkumu, fázi VaVaI (životního cyklu výzkumného projektu), formálních podmínkách i neformálních zvyklostech a kultuře. V tomto prostředí probíhá kooperace, strategie, politika, finanční toky a kontrolní mechanismy, odpovědnost, hodnocení a audit. Paula Boddington (2023, s. 12) přirovnává současný stav, kdy společnost ovlivňují technologie obsahující/využívající UI, v tom, jak interagujeme se světem, mezi sebou i jakým způsobem získáváme informace a znalosti, k jízdě „na hřbetě draka".

Upozorňuje tím na to, že se jedná o tak komplexní situaci složenou z různých dějů, které probíhají jak paralelně, tak za neustálého vzájemného ovlivňování, že mění naše představy o tom, co se děje, zatímco se tak děje. „Snažíme se přemýšlet o něčem, co zároveň může měnit, jak o tom smýšlíme v momentě, kdy tak činíme." To je větší výzva než pouze řešit něco nového, co se rychle vyvíjí." V souvislosti s regulací UI na evropské úrovni (AI Act) se používá

ve veřejném prostoru výraz: „střílet na pohyblivý terč“. Nové technologie, vč. UI se vyvíjejí rychleji, než regulace a další kontrolní mechanismy.

2.1.16 Globální organizace, mezinárodní instituce

Pro zaručení bezpečného využití nových technologií a minimalizaci jejich negativního vlivu na současnou i budoucí společnost, vznikly, a stále vznikají, různé mezinárodní odborné poradní orgány, interdisciplinární expertní skupiny, poradenská centra i mnohé výzkumné laboratoře a oddělení, a to při stávajících mezinárodních i nadnárodních organizacích, jako například:

UNESCO – Vyvinulo nástroj na hodnocení etického impaktu (2023a) a příručku Doporučení v rámci etiky UI: klíčová fakta (2023b).

OSN (United Nations), OSN a jejich iniciativa AI for Good,

OECD.AI vydalo katalog nástrojů a metrik pro důvěryhodnou UI (2024). Ten se pilotně ověřuje v rámci provincie Fryslân, Nizozemsko (2023). Součástí je slovní analýza etiky a udržitelnosti.

Na národní úrovni jednotlivých států nebo přímo inter-/multidisciplinární výzkumná centra či instituty na univerzitách a dalších výzkumných institucích.

GPAI, the Global Partnership on AI, Dutch AI Alliance (ALLAI-NL), the World Economic Forum Council on AI, and the High-Level Expert Group on the Implementation of the UNESCO AI Ethics Recommendation . European Commission High Level Expert Group on Artificial Intelligence, the IEEE Global Initiative on Ethically Aligned Design of Autonomous and Intelligent Systems, the Delft Design for Values Institute, the European Global Forum on AI (AI4People), the Responsible Robotics Foundation, the ADA-AI foundation, the European Artificial Intelligence Association (EURAI).

Globální partnerství UI (GPAI) je iniciativa partnerů z řad občanského sektoru, výzkumných organizací a akademie, vládních organizací a průmyslu, která sídlí při OECD.AI. Jejím cílem je přemostit prostor mezi teorií a praxí UI

podporou mezinárodního špičkového výzkumu a aplikovaných aktivit zaměřených na priority UI (GPAI, 2024).

Tento výčet není rozhodně úplný. Jedná se převážně o významné nadnárodní či evropské instituty a nejsou zde vyjmenována další univerzitní pracoviště, centra a interdisciplinární laboratoře. Přesto tento výčet naznačuje, že ekosystém organizací, které se etikou nových technologií a zejména UI zabývají, je velice složitý Přehledovou studii publikovali Ho et al. (2023, s. 1-19), ve které shrnuje cíle a činnost/poslání těchto organizací.

Institucí je mnoho. Někteří vědci však působí v rámci několika institucí či hodnotících komisí a poradenských výborů. Dochází tak k jejich síťování a přenosu informací, pokud se nejedná o tajný výzkum, mezi jednotlivými pracovišti a výzkumnými týmy. Vzniká však velmi mnoho reportů, publikací a vědeckých článků.

2.1.17 Municipality

Implementují řešení „Smart cities“ (SMOCR, 2020), ale i transparentní místní správa. Např. **Amsterdam** nebo **Helsinki** mají rejstřík algoritmů,[53] které město využívá, do kterého může kdokoliv nahlédnout – zajištění transparentnosti, férovosti. Občané mohou také dát zpětnou vazbu, sdílet své názory. Je to způsob participativního vývoje algoritmů, které jsou „zaměřené na člověka“ (human-centered algorithms). Rejstřík je průběžně aktualizován.

Město **New York** přijal místní zákon na ochranu žadatelů o práci před diskriminací způsobenou algoritmickým biasem (YC LL 144-21) (Deloitte, 2023).

2.1.18 Firmy, tržní analýzy a prognózy

Analýzy, indexy a prognózy dalšího vývoje založené na datové analýze. V podkapitole 1.3 byl již zmíněný „AI Index“ „Stanfordského institutu pro UI zaměřenou na člověka“.[54] Dalšími jsou např.: IBM Global AI Adoption Index 2023

[53] https://algoritmeregister.amsterdam.nl/en/ai-register/, https://ai.hel.fi/en/ai-register/
[54] Za rok 2023 je index k dispozici zde: https://aiindex.stanford.edu/report/.

Zajímavý report je Index připravenosti vlád na UI, kterou vypracoval Oxford Insights (2024). Do analýzy bylo zahrnuto 193 zemí, byly srovnávány v 10 dimenzích s 39 indikátory, rozdělenými do 3 sloupců: vláda, tech-sektor, data a infrastruktura). Česká republika je na 31. místě za rok 2023.

Indexy umožňují meziroční srovnání vývoje v oblasti UI, sledování nových trendů a odhad jejich budoucího vývoje.

Firmy - Technologie založené na UI fungují podle Jennifer Cobbe a Jatinder Singh (2023, s. 1189)[55] v rámci dodavatelsko-odběratelských řetězců, které jsou založené na datech. Zároveň jsou závislé na různých externích službách (jako cloudové služby, datových centrech, zdrojích dat od třetích stran, serverech, zdrojových obsahových sítích, UI technologiích poskytovaných formou služby technologickými giganty jako Amazon, Google, IMB atd). Tito aktéři, kteří často nevědí o všech ostatních v daném řetězci, jsou propojeni daty, která protékají od jednoho k druhému všemi směry.

Autoři popisují problematiku algoritmických dodavatelsko-odběratelských řetězců za účelem popisu jejich socio-technického kontextu, který se navíc neustále vyvíjí a mění (Cobbe, J., Veale, M. & Singh J., 2023, s. 1193). Na tomto kontextu závisí odpovědnost jednotlivých aktérů v rámci řetězce za zodpovědný přístup při vývoji, ověřování a nasazení algorimů v praxi. Tito aktéři si však i dané řetězce, podle autorů, (tzv. horizont odpovědnosti)[56] upravují tak, aby tuto odpovědnost minimalizovali a zároveň maximalizovali obchodní profit.

---

[55] Vědci působí v rámci multidisciplinární skupiny Compliant and Accountable Systems na Katedře počítačů Univerzity v Cambridge. Cobbe, J., Veale, M. & Singh, J., 2023. Understanding accountability in algorithmic supply chains ACM International Conference Proceeding Series, Doi: 10.1145/3593013.3594073 (1186 – 1197).

[56] Tzv. Accountability horizon potom omezuje visibilitu (orientaci) skrz tyto odběratelsko-dodavatelské řetězce a toho jednotliví aktéři využívají ke skrývání rozsahu své skutečné zodpovědnosti a minimalizují tak možnou detekci a postih svého chování (Cobbe, J., Veale, M. & Singh J., 2023, s. 1193-1194). Tento problém se „v kontextu literatury o algoritmické odpovědnosti nazývá problémem ´mnoha rukou´.

2.1.19 Univerzitní aliance

Další možností a příležitostí pro spolupráci univerzit na velkých společenských a civilizačních tématech, kam patří AI etika i udržitelnost, je jejich zapojení do tzv. univerzitních aliancí – příkladem je Aurora nebo EuroTeq pro technické univerzity.

2.1.20 Adaptace

Adaptace na nové podmínky a přijetí nových postupů a technologií – např. s ohledem na čím dál větší využívání generativní UI v kontextu mnoha profesí, přijala Evropská Komise tuto technologii a vydala „Živoucí" pravidla pro odpovědné využívání generativní UI ve vědeckém výzkumu (EC, 2024, březen). Počítá s tím, že se budou postupně upravovat podle dalšího technologického vývoje. Je to opět malý střípek do mozaiky všech činností, které se v ekosystému VaVaI dějí.

2.1.21 Interdisciplinarita

Interdisciplinarita jako mechanismus – interdisciplinarita je důležitý mechanismus. Umožňuje vědcům nacházet řešení složitých vědeckých otázek. Zlepšuje jejich schopnosti složité otázky si nejprve pokládat nebo je pokládat vědcům jiných oborů, porozumět kontextu své práce z jiného úhlu pohledu, včetně jeho dopadu, vidět nové výzvy a příležitosti a dělat lepší vědu. Podle Vantard, Gallard & Knoop (2023, s. 711) nelze posuzovat výsledky interdisciplinárního výzkumu pouze skrze optiku „součtu přínosu individuálních oborů."

Jak uvádějí, největším přínosem je to, co se děje na rozhraní těchto oborů (v tzv. interface), když se vzájemně tyto znalosti propojí a interagují. Efekt přínosu interdisciplinárního přístupu se potom násobí, respektive vědění se posouvá do prostoru/dimenze, které předtím nebyly žádným z individuálních oborů pokryty. Kombinace různých oborů (a dimenzí – příležitostí, které se takto mohou otevřít a pomoci nalézt řešení) je prakticky nekonečná. Na své přednášce

Filosofie a věda, hovořil Filip Grygar (6. 4. 2023) o přínosu filosofie, filosofického pohledu v kombinaci s fyzikou, který prokázal na několika příkladech z historie tohoto oboru.

Výzvou pro vědeckou kariéru ale potom je např. publikování a objektivní hodnocení jejich výsledků (např. formou peer review), protože je většinou provádějí zástupci individuálních tradičních vědních disciplín, kteří tento unikátní kombinovaný vhled sami nemají.

Podmínka interdisciplinarity nebo interdisciplinarita jako mechanismus ovlivňující kvalitu a bezpečnost vyvíjených systémů postavených na nebo využívajících UI, tak, aby byla zajištěna nejen její bezpečnost (jako minimální etická podmínka „Do no harm.“), ale aby byla přínosná a mohla být využita jako nástroj k řešení globálních problémů lidstva (viz SDGs definovaných OSN v předchozích bodech).

Aktivní podporu interdisciplinarity zajišťují některé výše zmiňované profesní organizace, které organizují interdisciplinární konference. V České republice je to tradiční („československá“) konference „Kognice a umělý život“, kterou spolorganizují technické univerzity a Centrum Karla Čapka, organizace SEFI (Evropská asociace pro vzdělávání inženýrských oborů) pořádá každoročně velkou konferenci pro své členy s širokým spektrem témat, která se vážou k podpoře rozvoje inženýrského studia (všech oborů STEM), např. v oblasti udržitelnosti, diversity, etiky a nových pedagogických přístupů k výuce.

Profesní organizace ACM pořádá interdisciplinární konferenci FaccT, kde podporuje příspěvky od akademiků i praktiků, které zajímá „férovost, odpovědnost a transparentnost socio-technických systémů (ACM, 2024).

2.1.21.1 Interdisciplinarita jako spolupráce

vědců a vědkyň s různou oborovou expertizou ve výzkumné organizaci, univerzitě, např. ve formě dočasného výzkumného projektu či konsorcia (financovaného např. z HE), které po určité časové období řeší konkrétní

výzkumné téma/výzkumnou otázku, problém, zadání atd. Je zde potřeba splnit podmínku finanční. Koordinace těchto projektů a efektivní komunikace jsou základem úspěchu.

Zde může být důležitá i osobnost manažera projektu, který zajišťuje hladkou organizaci, komunikaci a propojování vědců a vědkyň z různých vědeckých disciplin, zvyklých na jinou kulturu práce např. ze své „mateřské" organizace a zároveň jsou i z jiných zemí a kultur. Výzev, jak co nejefektivněji využít jejich znalosti a co nejrychleji tyto časově omezené projekty nastartovat, aby směřovaly k efektivní spolupráci a plnění slíbených „indikátorů a deliverables", je zde mnoho.

Rozhovor s PR manažerkou projektu HE, MM z Itálie (6. 3. 2024):[57] ...

„...Projektové manažerky, co jsem s něma v kontaktu, jsou totálně přetížené a vůbec nejsou schopné dělat cokoliv "navíc" v těch projektech. Je to škoda. Ty výsledky tam jsou, ale vědcům stačí, že si napíšou papery. Za ty peníze ... by ty projekty mohly mít mnohem vyšší impakt."

Problematika je zde ještě jiná. Projektoví manažeři mají na projektech často velmi malý úvazek (někdy žádný) a dělají jich souběžně mnoho. Projekty nefinancují někoho, kdo by cíleně interdisciplinaritu a maximalizaci jejich impaktu koordinoval (tzn. aby na to měl skutečně zaplacenou časovou kapacitu). Mohlo by se jednat i o různé odborníky na jednotlivé projektové fáze, ti co by podpořili zahájení spolupráce – už např. při psaní grantového proposalu, potom v rámci kick-off mítingu a následujících projektových fází, v jednotlivých projektových balíčcích, až po zajištění co nejvyššího konečného impaktu a diseminace výsledků, vč. případné koordinace s projekty ze stejného clusteru, které řeší stejné téma (v tomto případě „odpovědnou" UI).

V projektech je těžké předvídat, jestli stanovené indikátory vůbec splní, jestli jsou dobře nastavené. Pokud se ale podaří přinést opravdu zajímavé

[57] Překlad vlastní, autorka.

výsledky, není v kapacitě projektového týmu (a v individuálním zájmu participujících vědců) zajistit co nejefektivnější využití těchto výsledků nad rámec stanovených cílů.

2.1.21.2 Interdisciplinární výzkumná centra

Mohou být dalším stupněm předchozí spolupráce, která se **institucionalizuje** – např. v rámci udržitelnosti projektu. Nebo pokud projekt cíleně sloužil k zafinancování počáteční, rizikové/kritické, fáze fungování této nové organizace, jež v průběhu projektu ověřila postupy spolupráce, interní procesy, mechanismy, zajistila postupné vybudování renomé díky špičkovým vědcům, kteří jsou v jejich čele a výsledkům, které získala ještě v „projektové" fázi. atd.

Na tomto principu fungují projekty financované Evropskou komisí prostřednictvím Evropské agentury pro výzkum (REA, 2024) ERA Chairs , atd. V Praze tak vzniklo na Filosofickém ústavu AV ČR v roce 2023 nové centrum CETE-P, které je zaměřeno na výzkum, kde se protíná etika v oblasti enviromentální a technologické.

Etika a filosofie v kombinaci s přírodními vědami, informatikou, UI, to jsou centra, na která jsem se v rámci této práce zaměřila. V roce 2018 bylo založeno Centrum Karla Čapka pro studium hodnot ve vědě a technice, které vzniklo jako mezioborová platforma Ústavu státu a práva, Filozofického ústavu a Ústavu informatiky AV ČR a Přírodovědecké fakulty UK. Jejich cílem je spolupracovat na grantech i v rámci Strategie AV 21 a působit v rovině filosofické, ale i na legislativu. V Bratislavě sídlí interdisciplinární Kempelen Institut.

Zabývají se vývojem a přijetím mezinárodních iniciativ v oblasti politiky a strategických pokynů pro výzkum a aplikaci umělé inteligence. Jsou také autory definic a principů v oblasti etiky UI. Podobně vznikají i různé odborné výzkumné a poradní orgány, zájmové skupiny, organizace a podpora zodpovědného

výzkumu v oblasti umělé inteligence, např. (jedná se pouze o ty, které sleduji či přímo cituji v této práci, ne o úplný výčet):

The Alan Turing Institute, Responsible AI Institute, All Tech is Human, The Institute for Ethical AI and Machine Learning, Institute for Ethics in Artificial intelligence – TUM, Stanford Institute for Human-Centered Artificial Intelligence, Center for Humane Technology, Data and Society Research Institute, Centre for Data Ethics and Innovation, The Future Society, Ada Lovelace Institute, Center for Democracy and Technology, Berkman Klein Center for Internet and Society na Harvardově Univerzitě, Partnership on AI, AlgorithmWatch, Montreal AI Ethics Institue, The Algorithmic Justice Leage, Future of Life Institue (FLI), Robotics and AI Law society (RAILS), Digital Legal Lab, Center for AI and Digital Policy, The Centre for Human-Inspired Artificial Intelligence (CHIA) na Univerzitě v Cambridge, Institut pro Globální Priority se sídlem na Univerzitě v Oxfordu či tamtéž sídlící Institut Budoucnosti Lidstva, kde působil i český vědec Jan Kulveit (viz část 1.1.5 Interdisciplinarita), a další instituty, centra a interdisciplinární pracoviště.

2.1.21.3 „Skutečná“ interdisciplinarita

tak, jak ji definoval Jan Kulveit – viz 1.1.5. Základní myšlenkou je naučit se aspoň částečně podstatu jiného oboru, který potřebuji k řešení své vědecké otázky. Mezioborová interakce (interface), ta výše popsaná nová dimenze, se potom prakticky otevírá přímo v našem mozku. Z tohoto pohledu je přístup popsaný v 2.2.21.1 málo efektivní a vyplývá to i s rozhovorů s vědci. Např. ukázka z rozhovoru s TK z 9. 11. 2022:

LS: „*Hm. Když ta game theory vychází z té teorie, z té ekonomie, sociologie, matematiky. Byl některý z těch projektů, na kterých jsi pracoval vyloženě interdisciplinární? V tom smyslu, že tam byli lidi z těch ostatních oborů?“*

TK: *„Dělali jsme, měli jsme takový grant. Dokonce víc takových grantů, kde byli lidi z filosofie dokonce nebo z logiky, kteří se na tohle zase koukají trochu jinak. Ano, z těchto oblastí. Takže kombinace informatiků se zájmem o game theory, logiku, filozofové. Se sociology jsem nepracoval, ani moc se sociálními vědci, to ne. Ale s lidma s tímto společenskovědním vzděláním a snahou o to modelovat nějaké společenské nebo informatické interakce, to ano.“*

LS: *„Jak se ti pracovalo v tom týmu (pozn. interdisciplinárním)? Jak fungovala ta komunikace?“*

TK: *„Nefungovala.“* (smích)

TK: *„To by se nám asi nikdy nepovedlo nějak ty naše pohledy na věci sjednotit. Myslím si, moje zkušenost s těma, tzv. disciplinárníma týmama je dost špatná, no. V podstatě si tam každý dělal něco svého. Málokdy to mělo nějaký přesah v tom smyslu, že jeden by si řekl: ‚Naučím se něco od toho druhého.' Moc ne. Vždycky to byla taková kombinace jenom věcí.. Z toho jsem neměl moc dobrý pocit a byl to vlastně jeden z důvodů, proč jsem přešel sem.“*

2.1.21.4 Vznik nových interdisciplinárních oborů

Ze „skutečné“ interdisciplinarity potom mohou vznikat úplně nové obory. Příkladem může být sociální a humanoidní robotika, neurovědy, bioinformatika a umělá inteligence, jako taková.

Tato část se tedy snažila odpovědět na výzkumnou otázku: Jaké kontrolní mechanismy ovlivňují výzkum a výstupy v projektech využívajících nebo vyvíjejících umělou inteligenci (AI), aby byly v souladu s etickými principy?

Interdisciplinarita je hlavním tématem, které se touto prací prolíná a z textu vyplývá, že je vlastně sama důležitým mechanismem, který ovlivňuje kvalitu systémů využívajíců UI a zajišťuje předpoklad, aby tyto systémy byly i etické a společensky prospěšné, nebo aby minimálně nepůsobily negativně – viz výzvy a hrozby spojené s UI v části 1.3. Jaké jsou tyto etické principy UI a jaké jsou

podmínky tzv. vědecké integrity a proč je důležitá? Tím se zabývá následující kapitola.

3 Odpovědný výzkum a inovace (RRI)

„Etika UI je soubor hodnot, principů a technik, které využívají široce uznávané zásady toho, co je správné a co špatné, aby řídily morální chování při vývoji a používání technologií UI (Leslie, D., 2019, s. 3).“

Podle Pauly Boddington (2023, p. 22) je ke správnému použití etiky v praxi nezbytné „detailní porozumění kontextu, relevanci empirických dat, metod pro interpretaci komplexních sociálních fenoménů a různé způsoby porozumění lidským bytostem.“ Boddington (2023, s. 12) upozorňuje, že v souvislosti s UI si musíme „pokládat hluboké otázky nejen o technologiích, jejich přínosech a hrozbách, ale o našich lidských hodnotách - o nás.“ S tím potom, podle ní, souvisejí všechny obavy, naděje a polemiky, které v rámci etiky UI existují.

K tomu, abychom si mohli odpovědět na výzkumnou otázku: Jaké kontrolní mechanismy ovlivňují výzkum a výstupy v projektech využívajících nebo vyvíjejících umělou inteligenci, aby byly v souladu s etickými principy?, je potřeba si v rámci této kapitoly doplnit právě tuto etickou dimenzi, kam patří pojmy jako hodnoty, morálka a integrita. Jak tyto nehmotné kvality propojit s technickým světem algoritmů, automatizace, robotů, technologií, virtuální reality, augmentace a efektivity?

Mezioborová spolupráce a interdisciplinarita vědců je prakticky již nutným předpokladem, aby byly systémy, které využívají UI, vyvinuty a nasazeny v **souladu s lidskými hodnotami a cíli**. To je oblast, která se nazývá **AI Alignment**, tedy slaďování. Proces, který se snaží na lidské hodnoty a cíle nejen brát ohled, ale centrálně je zakódovat přímo do UI systému s využitím různých postupů a technik. Kvalitní, bezpečná a ověřená trénovací data ve strojovém učení, jsou

důležitým předpokladem, že i výstupy budou odpovídat těmto hodnotám. Kapitola se zabývá jednotlivými principy, jako hodnotami a znaky etické UI a problémy, které mohou nastat.

Při využívání neuronových sítí např. vzniká fenomén tzv. „**černé skříňky**". Tedy jedná se o nevysvětlitelný UI model, kdy neumíme přesně určit, jak dospěl ke konkrétním výsledkům. Problémů spojených s UI a jejímu nasazení prakticky ve všech oblastech lidského života, je mnoho a rozsah této práce ani expertiza autorky nedovolují se jim po technické stránce podrobně věnovat.

Tlak a nároky nové technologie kladou na všechny profese, nejen na **zodpovědnost a expertizu** jejich tvůrců. Většina profesionálů čelí již nyní výzvám nebo v budocnosti bude řešit, jak s těmito technologiemi efektivně a bezpečně pracovat. Kontinuální **vzdělávání** bude potřeba prakticky u všech profesí. Političtí lídři musí obstát ve své **regulační roli** (**AI Governance**), ale musí být schopni predikovat i budoucí vývoj, aby dobře nastavili strategické cíle v oblasti **digitální agendy** a podmínky pro bezpečné a férové využití UI.

3.1 Vymezení problematiky, základních pojmů a souvislostí

Klíčové pro úspěšnou implementaci **principů etické UI** do běžné praxe je, aby jejich funkce nebyla pouze **formální** – tedy kontrolní, „legislativní" (z pohledu mnohých vědců a převážně byznysu „omezující"), ale převážně **praktická** a **kreativní**. Taková, která vede ke kvalitnějším výsledkům v oblasti vědecké excelence, a v byznysové sféře ke kvalitnějším a bezpečnějším technologickým řešením a produktům, které budou úspěšné na trhu. K tomu mají sloužit definované principy jako nejvyšší forma hodnot, ke kterým výzkum a své cíle směřovat - otázka: **Co?**

Na operativní úrovni to potom budou sety otázek a následujících opatření, které z těchto hodnot budou vycházet a zároveň budou reflektovat účel, využití, typ výzkumu, cílové skupiny a jejich zájmy, a další. To si můžeme představit ve

formě "prováděcí vyhlášky" nebo praktického návodu, který nám dá odpověď na otázku: **Jak?**

Zatímco k principům etického UI vývoje je možné se veřejně přihlásit,[58] deklarovat je např. v **etickém kodexu** organizace, PR a další komunikaci, jejich naplňování a interní prostoupení těchto hodnot do všech procesů organizace, a nezáleží až tak, jestli se jedná o výzkumnou organizaci nebo korporaci, je mnohem složitější interně prosadit a implementovat, pokud zde není jasná motivace a interní koordinace.

Slovy Isidorose Karatzase (MUNI, 2023, 21. 11.), vedoucího sektoru pro výzkumnou etiku a integritu, na Generálním ředitelství pro výzkum a inovace Evropské komise:

„Důvěra je vše. Trvá roky, než se vybuduje, ve vteřině ji můžeme zklamat a náprava trvá věčně."

Etickými se staneme jedině tak, že se budeme **chovat** v souladu s etickými zásadami. Jejich deklarace v kodexu organizace tedy určitě nestačí. Jakmile si je však osvojíme, fungují jako prevence, pojistka proti „požáru", jak sám uvedl. Klíčové je, si tyto zásady uvést do každodenní praxe ve všech fázích výzkumného procesu. V oblasti UI však nejsou tak přesně definovány a nemají takovou tradici, jako například bioetika. Proto přiznal, že i na straně regulátora mají s dohledem nad agendou v oblasti UI etiky zatím nedostatek odborníků a zkušeností.

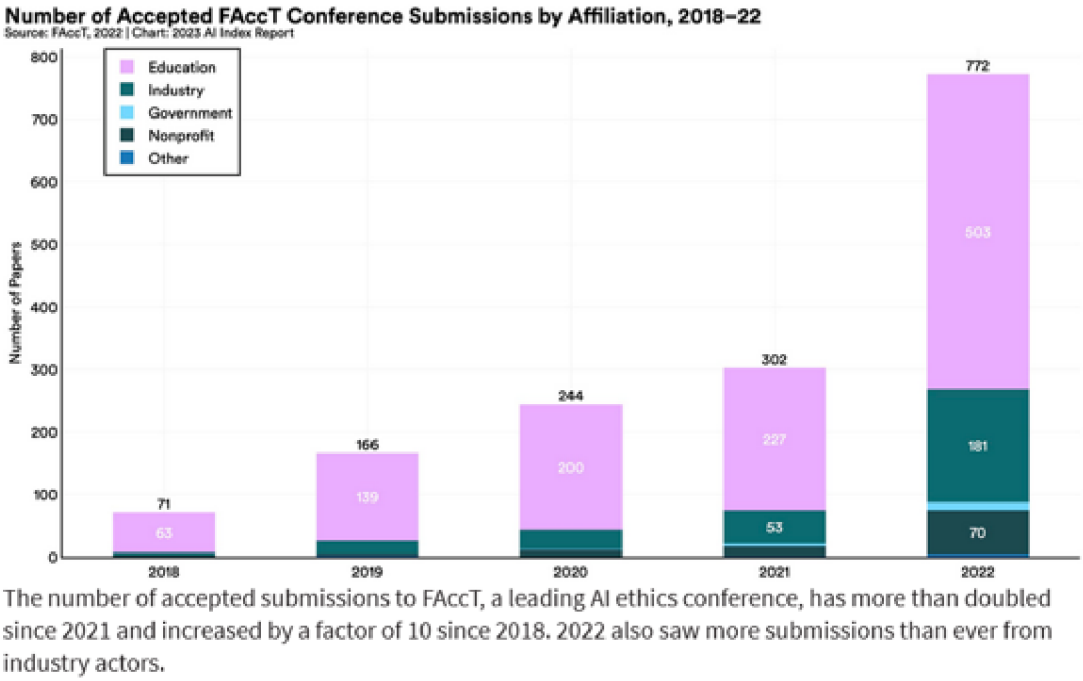
Zároveň mají evidenci, že úspěšné (v projektech HE a čerpání evropských fondů) jsou takové výzkumné organizace, které mají velmi dobře zpracovanou a internalizovanou agendu etiky a vědecké integrity, jsou potom pro EU komisi důvěryhodnými partnery, a ty, ve kterých dobře funguje administrativní a

[58] Za velký úspěch administrativy Biden-Harrisové se považovalo to, že se v červenci 2023 dobrovolně přihlásilo k plnění principů „zodpovědného vývoje" UI sedm největších amerických korporací : Amazon, Anthropic, Google, Inflection, Meta, Microsoft, a OpenAI. Tyto korporace se zavázaly mimo jiné také k prioritizaci výzkumu společenského dopadu jimi vyvíjených technologií využívajících UI (White House, 2023).

manažerská podpora vědců (RMAs). Větší roli by v zajištění etiky a výzkumné integrity měly podle Karatzase hrát výzkumné instituce a vědce aktivně podporovat. Nejen při přípravě grantové přihlášky, ale v zajištění a zjednodušení všech interních procesů, vzdělávání a další podpoře. V současné době tato agenda připadá většinou na vědce samotné.

Hypotéza: Na začátku výzkumu jsem vycházela z předpokladu, že propojení oborů UI a etiky prakticky zcela chybí, protože se nepropisuje do vědecké praxe managementu projektů tam, kde je UI předmětem výzkumu nebo kde je využívána jako metoda/nástroj k dosažení cílů v jakémkoliv jiném oboru nebo praxi.

Zjištění: Etika UI je prudce se rozvíjející obor. Zájem o téma etiky umělé inteligence můžeme sledovat nejen v rostoucím počtu odborných publikací, ale roste i počet vědeckých článků přijatých na největší konferenci s touto tematikou FAccT, kterou pořádá Asociace pro výpočetní techniku (ACM – Association for Computing Machinery). Literatury a zdrojů je tedy čím dál víc, do praxe se však její výstupy a zjištění nepropisují tak rychle. Hypotéza se nepotvrdila.



Obrázek 8 ***Počet rostoucího počtu submitovaných článků na interdisciplinární konferenci FaccT s tematikou etiky UI*** **(ACM, 2023).**

Kromě zkušeností s projekty zaměřenými na výzkum a vývoj UI, kde je, ze strany poskytovatelů grantů kladen čím dál větší důraz na zhodnocení etických aspektů a impaktu, ale na úrovni bakalářského a magisterského studia technických oborů v ČR je v tomto ohledu naprosto zanedbatelná nabídka kurzů ze strany technických vysokých škol, našla jsem v literatuře práce až 30 let staré, které volaly po větším zohlednění etických aspektů ve výzkumné praxi technických/inženýrských oborů.

Matthew Charlesworth provedl velmi pečlivou analýzu principů a kurzů počítačové/informační etiky již v roce 2005.[59] Práce obsahuje přehlednou historii vývoje oboru „počítačové etiky" (Computer Ethics), který položil základy dnešní etice UI i s úvahami o negativním dopadu nových technologií na společnost a např. přístupu v oblasti senzitizace vědců i veřejnosti a využívání případových studií (v oboru se tehdy profiloval prof. Floridi, který v něm působí dodnes).

[59] Nejedná se o etiku UI, protože v roce 2005 tento obor nebyl jestě tak populární – tzv. "UI zima".

Konkrétní problémy, které řešila Computer Ethics, mají však málo společného s dnešní problematiou etiky UI už vzhledem k rozsahu společenského dopadu těchto technologií. Dříve byl negativní efekt více lokalizován, tzn. byl malého rozsahu, ale i snáze dohledatelný. V důsledku internetu, cloudových řešení, velkých objemů dat dostupných ve zlomku vteřiny z kteréhokoliv místa na Zemi, které je na internet napojené, mají důsledky možného neetického přístupu velký impakt, který není vždy možné předvídat. V rámci etických sebehodnocení ve výzkumu je však potřeba tyto dopady promyslet nebo vyhledat pomoc odborníků v daném oboru, vč. oborů SSH, pokud na to nestačí expertiza daného hlavního řešitele.

Na druhou stranu, s přihlédnutím k této dlouhé historii oboru etika UI, pokud tam počítáme i počítačovou etiku (Computer Ethics) se občas zdá, že nejvíce diskutovaným etickým problémem ve spojení s UI, který rezonuje ve veřejném prostoru, je ryze filosofické tramvajové dilema, které bývá v různých alteracích zmiňováno v kontextu problematiky autonomních vozidel.[60] Je potřeba etiku prezentovat jako součást každodenních rozhodnutí a problémů, které jsou už dnes relevantní, jako např. využívání systémů podporovaných UI ve veřejném prostoru (viz příklad 2.2.17 a veřejně dostupné rejstříky algoritmů, které využívají municipality v rámci jejich Smart City projektů).

Karatzas (MUNI, 2023, 21. 11.) si uvědomuje, že na vědce z technickým vzděláním mohou působit etické otázky nejasně a frustrovat je, že diskuse nenabízí jasné řešení. Není to obor, kde by se výsledek dal spočítat matematicky, ale je tam mnoho proměnných, které je potřeba zvážit. Pomáhá dívat se na určitý problém za použití různých makroetických teorií a vybalancovat různé důkazy,

[60] Tedy, má-li vozidlo přejet 5 osob na kolejích nebo se jim aktivně vyhnout (přehodit výhybku) a přejet „pouze“ jednu osobu na vedlejší koleji. Tato úloha má různé alternativy, kdy rozhodnutí ovlivňují různé podrobnosti o daných osobách – např. ta jedna osoba na koleji je matka řidiče tramvaje, atd. Rozhodování v podobných situacích si mohou lidé vyzkoušet zde https://www.moralmachine.net/ The Moral Machine. Téma tramvajového dilematu bylo diskutováno i s J. Hvoreckým v rámci rozhovoru o etických tématech a filosofických přístupech k UI 9. 1. 2023.

které ale mohou jít někdy přímo proti sobě – jako transparentnost a ochrana citlivých dat zákazníka (viz ukázka z rohovoru s JV z 8. 11. 2022). Výsledek bude vždy kompromisem.

JV v rozhovoru z 8. 11. 2022 takovou situaci popsal: ...

„...V realitě je potřeba postupovat v souladu s etickými principy při zadávání daného úkolu, kterého se má celou operací dosáhnout a ošetřit v souladu s etickými principy, jak budou využita vstupní data a co se s nimi bude dít po dosažení našeho cíle, když například výstup předáme zákazníkovi. Co se s těmito podkladovými daty bude dít. Na jednu stranu by měly být výstupy transparentní, tedy to, jak jich bylo dosaženo, na druhou stranu je potřeba určitá citlivá data anonymizovat a chránit.“

Charlesworth (2005, s. 14) nabízí přehled makroetických teorií. Na základě rozhovoru s J. Hvoreckým z 9. 1. 2023 z centra CETE-P a Filosofického ústavu AV ČR jsem se dozvěděla, že filosofové v souvislosti s technologickou etikou používají většinou 3 z nich:

1. **Deontologii** – věří v existenci univerzálních pravidel, zcela racionálních a nerozporuplných, která platí za všech okolností. V případě tramvajového dilematu by, podle Hvoreckého, rozhodnutí přejet byť jednoho člověka, místo pěti, znamenalo, že zabíjet se obecně může. Nebo pravidlo „Nezabiješ!“ také nemůže platit obecně a za všech situací, např. ve válečném konfliktu, v sebeobraně, atd. Naprogramovat podle nich robota by také nebylo vhodné, neuměl by se přizpůsobit situaci.
2. **Etiku cti** (Virtue ethics) – Ta předpokládá, že se člověk bude chovat ctnostně, přiměřeně dané situaci, v souladu s morálkou. Nejsou tam jasně specifikovatelná pravidla, která by se např. dala použít při programování robota.

3. **„Utilitární etiku/Utilitarianismus** – pracuje s představou, že existuje algoritmus počítání dobra a zla. Musíme si zvolit, co je tou jednotkou dobra a zla. Ty situace poměřujeme prostě tím, co nám vychází lépe. Jako menší zlo.“ (J. Hvorecký, 9. 1. 2023.)

Těmto přístupům se věnuje i Steen (2023, s. 65) a přidává ještě 4. přístup **konsekvencialismus**. Ten, jak název napovídá, při zvažování, co je správné a špatné, bere v úvahu možné důsledky/následky daného rozhodnutí pro vás nebo někoho jiného, koho se týká.

„Etika ‚vědy' bývá někdy chápána ve velmi zúženém smyslu jako etika vědecké profese, jako ‚etika ve vědě'. V tom případě vyplývají etické normy ze zásad samotného provozu vědecké práce a problémy se týkají záležitostí, jako je plagiátorství, dodržování autorských práv, snahy převádět v zájmu ekonomických zisků v podstatě vědecké ‚objevy' na patentovatelné ‚vynálezy' apod. K tomu se přidružují specifika z konkrétních oborů. Větší závažnosti nabývají otázky, které se týkají odpovědnosti vědců za jejich objevy a vynálezy: vůči komu má vědec morální odpovědnost za výsledky své práce“ (Drozenová, 2010, s. 13)?

Drozenová upozorňuje na dokument Morálně-etické aspekty výzkumu a vývoje, který „jasně vymezuje morální a etické požadavky na vědní oblast.“ Dokument (1999, s. 6) v sekci věnované etickým zásadám instituce zabývající se výzkumem a vývojem, přímo uvádí, že by tyto instituce: „měly přijmout vlastní etické kodexy, které vycházejí z obecných principů etiky vědecké práce, ale obsahují i specifická etická pravidla pro příslušnou oblast výzkumu či vývoje. Při zachování svobody vědeckého bádání by tyto instituce měly zavázat své členy k dodržování příslušných etických pravidel.“ Tedy např. zavázat vědce k dodržování interního institucionálního etického kodexu, který se stane součástí – přílohou – pracovní smlouvy. Dokument (1999, s. 8) zmiňuje i etické a morální aspekty obranného výzkumu.

Důsledky možných morálních a etických dopadů UI na společnost se velmi záhy po iniciaci oboru UI začal zabývat **Norbert Wiener** z MIT. Již v roce 1960 publikoval článek „Morální a technické souvislosti automatizace“ (Christian, 2020, s. 239-240). Zamýšlel se nad důsledky toho, co se stane, když se UI vymkne kontrole. Zdůrazňoval, že je nezbytné, aby si člověk zajistil možnost zasáhnout v tomto případě do procesu a vždy jasně specifikoval cíle a podmínky, za kterých se bude UI využívat (Wiener, 1960, 1355-1358).

Dřívější představa vědců, včetně **Turinga**, byla, že v případě nebezpečí vypnou proud, odpojí zařízení ze sítě a nežádoucí proces se zastaví. Bohužel tohle by bylo možné jen v případě, že si nesouladu mezi plánem a výstupy UI stihneme vůbec všimnout. Realističtější scénář však je ten, že jednotlivé operace, které UI provádí – např. bankovní či burzovní operace, probíhají ve zlomku vteřiny (v cloudu) a nelze je jednoduše lokalizovat. Není možné mít nad jednotlivými akcemi kontrolu, natož systém vypnout.

V oblasti robotiky se za zakladatele etického přístupu považuje **Gianmarco Veruggio**, který v roce 2006 navrhl přehled možných výzkumných přístupů k robotické etice. Od té doby nastaly další změny, zejména ve vztahu ke klimatické krizi a sociálním, ekologickým a politickým souvislostem, pokud jde o zdraví naší planety (EcoSocioBotics, 2022). V oblasti robotické etiky a etiky informatiky je známý také uruguayský filozof, akademický profesor **Rafael Capurro.**

Úvahy o tom, co je morální ve vztahu k vývoji UI, interakci člověka a UI[61] a možnými budoucími „právy“ a morálním statusem UI, nám umožňují nastavit zrcadlo:

- co znamená být člověkem,
- jak definovat, co je lidské,

[61] Některé vědecké práce rozvíjejí úvahy, zda lze tuto interakci nazvat kolaborací a roboty spolupracovníky (Evans, Robbins, & Bryson, 2023, s. 1-20).

- jaký vztah a cíle má člověk ve světě a ve vztahu k dalším živým bytostem a k technologiím, které vytváří.

Tyto filosofické úvahy jsou důležité i vzhledem k tomu, abychom si uvědomili a definovali, jakou společnost v současnosti, a především v budoucnosti, chceme, jaké hodnoty jako lidstvo zastáváme a zda jsme schopni se na nich dohodnout napříč kulturami a různými tradicemi. Z tohoto úhlu pohledu jsou definované globální cíle udržitelnosti (SDGs) na úrovni OSN neuvěřitelným úspěchem lidstva.

3.2 Principy etické UI

Vývoj pravidel a legislativy v nových, prudce se vyvíjejících oborech, se může inspirovat historií a milníky, které jiné obory, jako bioetika, učinily před několika desítkami let. Zároveň je dobré kriticky zhodnotit, zda tato opatření byla účinná a co je potřeba specificky ošetřit jinak. V oblasti UI existuje velké množství principů, kodexů a zásad na mnoha úrovních. Centrum Berkmana Kleina Harvardovy univerzity (Fjeld, J. et al., 2020) sestavilo přehledovou mapu (tehdy) platných „Principů odpovědné UI" ve tvaru barevného terče. Jednotlivé sety principů etické UI jsou zde srovnány podle toho, zda zahrnují klíčová témata: soukromí, odpovědnost, bezpečnost, transparentnost a vysvětlitelnost, férovost a prevenci diskriminace, lidský dohled nad technologiemi, profesní odpovědnost, podporu lidských hodnot a mezinárodních lidských práv. Tato klíčová témata tvoří jednotlivé barevné vrstvy terče.[62]

Ten je rozdělen na sektory a jednotlivé aktéry v rámci těchto sektorů, kteří dané principy publikovali nebo se k nim přihlásili. Jedná se o organizace (aktéry) neziskového sektoru, soukromého sektoru (byznys), vlády (vládní organizace),

[62] V angl. originále: Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, Human Control of Technology, Professional Responsibility, Promotion of Human Values, International Human Rights (Fjeld, J. et al., 2020).

mezivládní organizace a uskupení sdružující různorodé zájmy (multistakeholder) – viz aktéři výzkumného ekosystému UI (viz podkapitola 2.1.2). Celkem hodnotí a porovnává 35 „principů odpovědné UI“ (existujících do roku 2020). Pokud tedy vezmeme v úvahu 12 největších výzev spojených s UI, definovaných v podkapitole 1.2, mohly by spadat pod principy:

- **férovosti a prevence diskriminace** (výzvy: bias, zaměstnanost),
- **právo na soukromí** (výzva: soukromí – privacy je přímo jedním z principů),
- **transparentnost a vysvětlitelnost** (výzvy: zkreslení/ misreprezentace, fenomén černé skříňky)
- **profesní zodpovědnost** (výzva: otevřená zdrojová data, open source, práva vyplývající z duševního vlastnictví a copyright)
- **bezpečnost** (výzva: existenciální rizika)
- **podpora lidských hodnot** (výzva: zaměstnání)
- **podpora mezinárodních lidských práv** (výzva: existenciální rizika)

Výzvy: Přístup ke konkurenceschopnému **výpočetnímu výkonu, přístup k datům, otevřená zdrojová data** jsou předmětem konkurenčního boje a je otázka, za jakých podmínek firmy jako Google, Facebook atd. tuto kapacitu/data zpřístupňují výzkumným organizacím. Open source data, která vznikla v rámci výzkumných grantů financovaných EU mají podmínku být publikována v režimu open source (Google, 2024).[63]

Na té nejvyšší úrovni, zároveň však velice široce definované, jsou morální zásady, které korespondují s hodnotami dané civilizace/společnosti. Spoléhá se na náš cit, že se ve složité situaci zachováme podle těchto nejvyšších zásad, ale v kontextu různých kultur mohou nabývat trochu odlišného významu a vysvětlení (Whittlestone et al., 2019). Velké množství principů však podnítilo i kritiku, že

[63] Př.: Google (2024) má tzv. Data Commons nebo data z tzv. Earth observation, umožňují využívat environmentální data a letecké snímky, které mapují vhodné lokality pro využití solární energie - viz podkapitola 1.3.

pouze definice principů nezajistí, aby se přetavily v konkrétní cíle, pravidla a skutečnou změnu v oblasti praxe.

V evropském kontextu Luciano Floridi, filosof a etik UI (Floridi et al. 2018) předsedal expertní vědecké komisi AI4People, která stanovila pět základních a nejšířeji definovaných principů pro "Good AI Society": prospěšnost, zásada neškodit, spravedlnost, (respektování) autonomie a vysvětlitelnost (transparentnost). První čtyři z těchto principů vycházejí z lékařské etiky/bioetiky (Hanemaayer, 64-65).

Floridiho (Floridi et al., 2018, s. 695 a Floridi, 2023, s. 59) etické principy jsou syntézou šesti mezinárodních deklarací, které prosadili sami experti v oboru UI. Jedná se o:

- Asilomarské principy vývoje UI z roku 2017 (the Asilomar AI Principles), vytvořené známými mysliteli v oboru UI z akademické sféry i průmyslu – místo Asilomar bylo zvoleno symbolicky, protože právě zde byla v roce 1975 přijata pravidla pro zákaz na práce s rekombinantní DNA.
- Montrealská deklarace za zodpovědný vývoj umělé inteligence z roku 2017 (Montréal Declaration for a Responsible AI Development);
- Globální iniciativa IEEE k etice autonomních a inteligentních systémů z roku 2017 (the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2017);
- Prohlášení Evropské komise k UI, robotice a autonomním systémům z roku 2018 (the Statement on AI, Robotic and 'Autonomous' Systems)
- UI Komise Sněmovny Lordů ve Spojeném království z roku 2017 (the House of Lords Artificial Intelligence Committee in the U.K.)
- a Partnerství UI v San Francisku z rok 2018 (the Partnership on AI in San Francisco), které reprezentovalo akademiky, organizace občanské společnosti a firmy vyvíjející UI technologie.

Odborná skupina na vysoké úrovni pro UI (HLEG), zřízená Evropskou komisí v červnu 2018 (Floridi z nich vycházel, viz výše) vydala Etické pokyny

pro zajištění důvěryhodnosti UI. Navrhla **rámec pro důvěryhodnou UI**. Důvěryhodná UI má 3 složky (HLEG, 2019, 4), "které by měly být dodrženy v průběhu celého životního cyklu UI systému:" **a) legální** – respektující veškeré platné právní a správní předpisy, **b) etická** – zajišťovat dodržování etických zásad, **c) robustní** – z technického sociálního hlediska, "jelikož **i dobře míněné systémy UI mohou způsobit neúmyslnou újmu**."

Tyto 3 složky musí fungovat souběžně, i když dokument se převážně zabývá etikou a robustností UI a ambicí tohoto dokumentu bylo navrhnout nejen seznam zásad, ale i vodítko, jak jich dosáhnout, aby mohly být uplatňovány v praxi. Dokument je prakticky rozdělen tak, aby postupoval od nejabstraktnějších pokynů, po ty opravdu nejpraktičtější v poslední kapitole.

Skupina HLEG úzce spolupracovala s Evropskou skupinou pro etiku ve vědě a nových technologiích (EGE), což je poradní skupina Evropské komise a stavěla na závěrech velkých projektů SHERPA a SIENNA.[64] Komise na těchto základních dokumentech postavila praktická doporučení pro etické samohodnocení projektů, vč. pokynů pro hodnotitele, a dále pracovala na legislativní části, tedy nařízení EU v oblasti umělé inteligence (tzv. AI Act – viz 2.2.7) a dalších tzv. digital acts.[65] V souvislosti s právní ochranou lidských práv, které také může UI poškozovat, se využívá GDPR[66]. EU je jedním z nejvýznamnějších aktérů ovlivňujících ekosystém VaVaI UI v ČR.

---

[64] Projekty SHERPA (https://www.project-sherpa.eu) a SIENNA projects (https://www.sienna-project.eu). Projekt SIENNA řešil hlavně tři nové technologické oblasti UI, kde zkoumal jejich etické a socio-ekonomické dopady – lidská genomika, lidské "vylepšování" – human enhancement, a interakce lidí a strojů. Projekt SHERPA se zabýval etickými dimenzemi chytrých informačních systémů.

[65] Jedná se o další regulace v oblasti digitálních technologií, jednotného trhu – Digital Markets Act – DMA a Digital Services Act DSA a nakládání s daty (GDPR)

[66] Problematiokou GDPR a ochranou lidských práv ve vztahu k hrozbám UI se zabývali Šmuclerová, Král, Drchal, 2023a, 17. Informace o ukončeném projektu: https://prg.ai/projekty/ai-lidska-prava/.

Překážky v oblasti legislativy VaVaI a implementace kontrolních mechanismů v oblasti etiky ve výzkumných projektech v rámci ČR[67] popsala prof. Veselská v rozhovoru pro magazín Universitas (Marušáková, 2023).

Prof. Veselská zmiňuje, že v České republice chybí:

"zákon o výzkumu, který by určoval obecná pravidla pro jeho provádění. Máme jen zákon o podpoře výzkumu, což je záležitost jeho financování z veřejných prostředků. ***Vůbec pak nemáme z právního hlediska komplexně ošetřené otázky etiky výzkumu na člověku****. Momentálně je u nás regulace etiky výzkumu na člověku řešena pouze ve dvou bodech:"*

1. *Pro akademické prostředí spíše okrajový, protože jde o povinné posuzování některých typů klinických studií etickými komisemi, což se týká zejména zdravotnických zařízení. I když kliniky fakultních nemocnic jsou společnými pracovišti s univerzitami a akademici na nich působí, pořád je to většinově záležitost resortu zdravotnictví, nikoliv školství.* ***Pro jakýkoliv další výzkum na člověku, který probíhá na univerzitách, české zákony posouzení etickou komisí nevyžadují.***
2. *Před více než dvěma lety ČR ratifikovala dodatek k Úmluvě o lidských právech a biomedicíně Rady Evropy o biomedicínském výzkumu.[68] Dokument je právně závazný a stanovuje pravidla pro povinné etické posouzení biomedicínského výzkumu. Není to ale nijak implementováno do českých zákonů* ***a opět se to týká jen biomedicínského výzkumu s tím, že v něm musí jít o nějakou intervenci na člověku. V ČR tedy není vůbec nijak systematicky ošetřeno dodržování etických standardů v psychologickém, behaviorálním či jiném typu výzkumu. V oblasti biomedicíny pak nemáme ošetřenou výzkumnou práci***

[67] Popisovaná situacie není specifická pro UI, ale pro celou oblast VaVaI v ČR. Lze tedy analogicky použít i pro UI Governance/Kontrolní a obranné mechanismy UI.

[68] Sdělení č. 30/2020 Sb. m. s. Ministerstva zahraničních věcí o sjednání Dodatkového protokolu k Úmluvě o lidských právech a biomedicíně souvisejícího s biomedicínským výzkumem, platné od 1. 9. 2020. Rozhovor je již z roku 2023 (14. 2.).

s pacientskými daty nebo biologickými vzorky tak, aby byla v souladu s mezinárodními etickými standardy."

Na otázku, jak je v projektech domácích grantových soutěží posuzována etika, aby byla v souladu s mezinárodními standardy, prof. Veselská dále odpověděla, že:

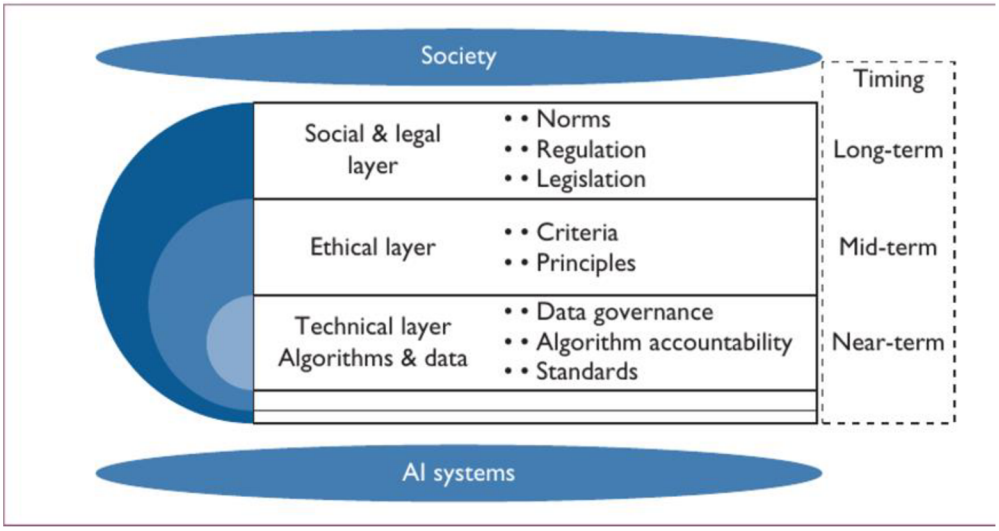
"*Hlavní problém je* ***u nás v nejednotném přístupu***. *U některých grantových agentur se otázka etiky neřeší vůbec – a tam, kde ano, tak bohužel nekonzistentně. Zadávací dokumentace jednotlivých soutěží většinou vyžadují schválení etické komise pouze tam, kde je to vyžadováno zákonem. Z mé předchozí odpovědi je ale jasné, že toto pojetí je nedostačující, protože stávající zákony ani zdaleka nepokrývají celou oblast výzkumu na člověku. Navíc není jasně stanovená posloupnost, v níž mají být posouzeny vědecké a etické aspekty výzkumného projektu. Jedné z našich grantových agentur stačí doložit schválení etickou komisí až tehdy, kdy je projekt vybrán k financování. Jiná grantová agentura naopak finální schválení projektu etickou komisí vyžaduje už ve chvíli, kdy se projekt podává do soutěže. Nicméně etická komise, pokud svoji práci dělá pořádně a v souladu s oborovými standardy, potřebuje na posouzení projektu určitý čas a většinou v této fázi nevidí projekt v jeho finální podobě. Odsouhlasuje tedy něco, co se ještě může změnit, aniž by o tom komise byla informována. Takové posuzování je v podstatě proti mezinárodně zavedeným standardům etiky výzkumu. Navíc komise věnují spoustu času zbytečně projektům, které v grantové soutěži neuspějí (Marušáková, 2023)*".[69]

V oblasti výzkumu UI ani zatím žádná legislativa platná není (AI Act se na výzkum UI nevztahuje – EC, 2024), nelze tedy čekat, že by vznikla podobná povinnost kontroly etickým výzkumným komisím např. na technických univerzitách, kde k výzkumu UI dochází. Pokud je však logika etické evaluace zaměřena pouze na konkrétní dopady na člověka, v případě UI by bylo velmi obtížné tento dopad posoudit ve fázi před podáním výzkumného projektu,

[69] Prof. Veselská má unikátní vhled a přehled o situaci v oblasti etiky a výzkumné integrity, odpověď jsem tedy nechala nezkrácenou (pozn. autorky).

pokud by se např. nejednalo o osobní data, nebo oblast algoritmu v oblasti bankovnictví či HR nebo přístupu k jiným službám, které by mohly způsobit diskriminaci, či interakci člověka a robota.

Dopady UI jsou ale komplexnější a často na velký segment společnosti, kdy může způsobit změnu chování (např. deepfake) voličů, atd. Jak tedy potom, aspoň teoreticky, hodnotit tyto možné dopady výzkumu ještě předtím, než bude výzkum zahájen (bude podána grantová přihláška) – viz komplexnost možných oblastí dopadu v závislosti na čase a jak je ošetřit v Obrázku č. 9.

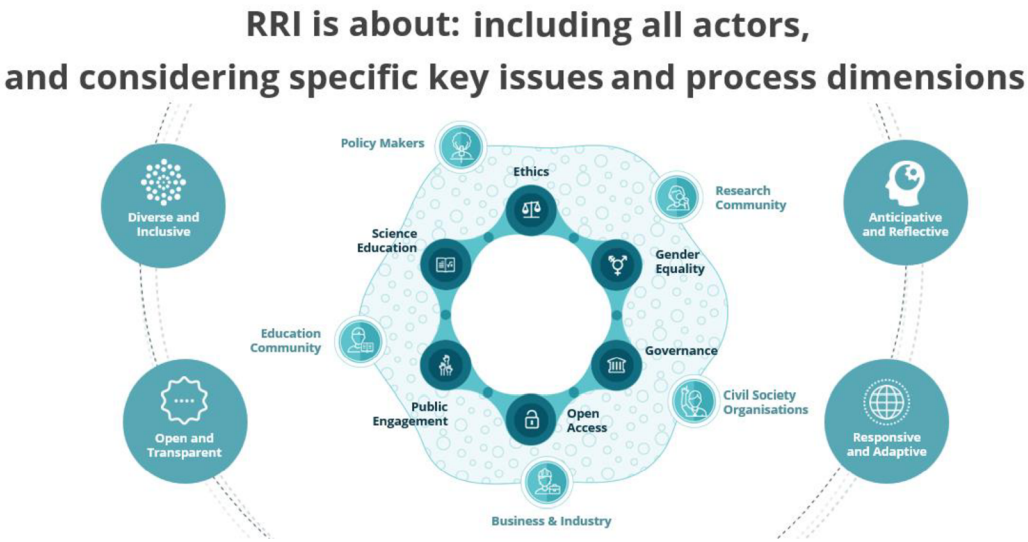


Obrázek 9– Model AI governance (autoři Gasser, U. & Almeida, V.A., 2017, s. 60).[70]

Jednotlivé vrstvy představují „filtry" opatření, která mají zajistit bezpečnost UI systémů pro společnost (minimalizovat jejich negativa). Bere v úvahu i faktor času a možný efekt dopadu UI systémů na společnost v závislosti na době jejich využívání. Kontrolní mechanismy (a zde se vracíme ke kapitole 2), které v rámci tohoto VaVaI UI ekosystému probíhají mezi jednotlivými aktéry za podpory formálních podmínek a procesů, jsou však součástí tzv. RRI (responsible research and innovation) – odpovědného přístupu k výzkumu a inovacím nebo se

[70] Gasser, U., & Almeida, V.A. (2017). A Layered Model for AI Governance. IEEE Internet Computing, 21 (6) (November): 58–62. doi:10.1109/mic.2017.4180835.

v literatuře užívá termín Socially Responsible Science (SRS) – společensky odpovědná věda. Etika a Governance jsou součástí ekosystému RRI stejně jako genderová rovnost, open access, zapojení veřejnosti/cílových skupin a vzdělávání. Ty jsou v následujícím Obrázek 10- Ekosystém RRI (responsible research and innovation) – odpovědného přístupu k výzkumu a inovacím[1] uvedeny do kontextu vědního ekosystému, kde jsou vyznačeni i aktéři a rozměry jednotlivých procesů mezi nimi jako: Variabilní a inkluzivní, otevřené a transparentní, Anticipativní a reflektivní, Responzivní a adaptivní.



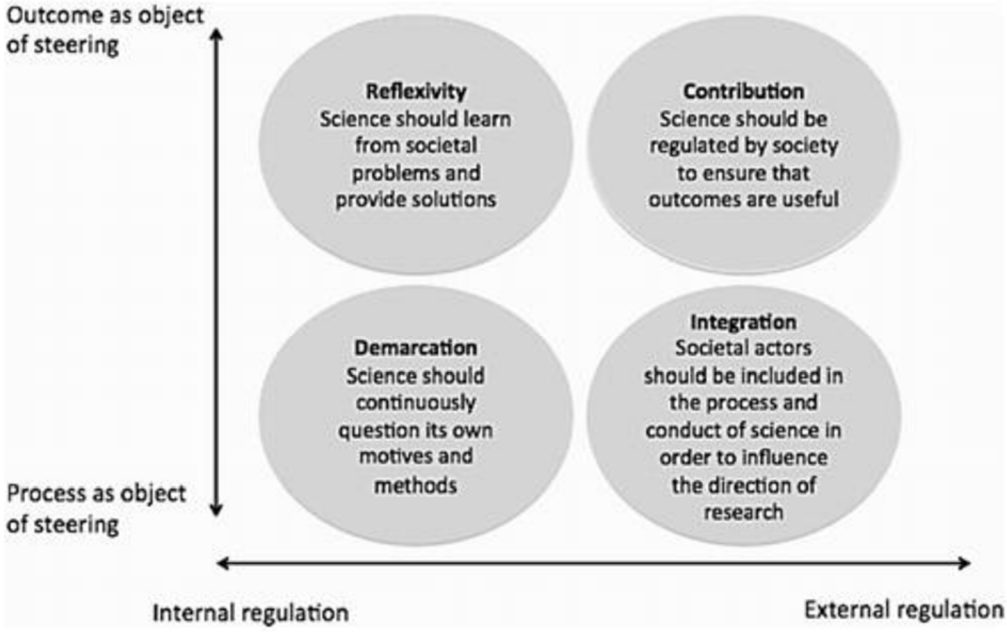
Obrázek 10- Ekosystém RRI (responsible research and innovation) – odpovědného přístupu k výzkumu a inovacím[1]

Resnik, D. B. a Elliott, K. C. (2015) uvádějí, že vědci se v rámci zodpovědného přístupu k výzkumu řeší tři základní okruhy etických dilemat:

1) dilemata ve vztahu k výběru vědecké otázky/vědeckého problému,
2) dilemata, která se vztahují k publikování výsledků a sdílení dat a
3) dilemata vztahující se ke spolupráci s cílovými skupinami/ obecně – společností.

Musí doslova „vybalancovat své profesní a společenské závazky a nezkompromitovat svou objektivitu."

Očekávání, která společnost (zadavatelé, poskytovatelé grantů, veřejnost) od vědců UI má a která tlumočí např. skrze podmínky financování jejich výzkumu,



Obrázek 11 ***Přehled 4 politických reacionalit uspořádaných v závislosti na tom, zda obhajují externí či interní regulaci vědy a zda-li navrhují, že je jejich sférou zájmu výzkumný proces či výsledky bádání (Glerup a Horst, 2014, s. 36).***

hodnocení a kontroly (společenské odpovědnosti, etického přístupu, odpovědnosti za výsledky) nebo prostřednictvím formulace strategických oblastí výzkumu, by se dala shrnout v rámci tzv. **politických racionalit**, jak je ve svém článku označily Glerup & Horst, (2014, s. 38-39).

Ze čtyř přístupů naznačených v obr. 11 lze k popisu současného stavu VaVaI ve vztahu k výzkumu nových technologií, vč. UI, použít politickou racionalitu, kterou autorky pojmenovaly „**Reflexivita**". Vyznačuje se zaměřením na

hodnocení výsledků/výstupů a jejich společenského dopadu s důrazem na interní „samoregulaci" vědce. Ten si má sám uvědomovat možné hrozby a negativní dopad svého výzkumu a být zodpovědný (s.38). Očekává se, že se bude v daných socio-ekonomických dopadech na společnost vzdělávat (s. 39), což popisuji dále v podkapitole 3.5. Aby byli vědci považováni za společensky odpovědné, což by se mělo projevovat jako forma sebe-uvědomění, schopnosti předvídat důsledky jejich vlastní praxe. „Dokud tak neučiní, není možné vědu považovat za společensky odpovědnou." Tento přístup je nyní prakticky preferován v grantových žádostech HE či v publikačních pravidlech některých publikačních domů a vydavatelství.

Není to zákon, ale je možné použít motivačního mechanismu (grant, financování výzkumu, publikační výsledky za které je vědec hodnocen), aby se změnil přístup vědce i v případě, že by se jinak o etický rozměr a impakt svého výzkumu nezajímal. Druhá varianta jsou vědci, kteří tímto způsobem již dávno přemýšlejí a postup etického sebehodnocení si již sami dobrovolně internalizovali.

UI se však týká i racionalita „**Přispění/umožnění**", že by věda měla přinášet inovace a znalosti a přispívat tak k národnímu a regionálnímu rozvoji a konkurenceschopnosti, tedy praktické cíle. Ne vycházet z čisté zvědavosti a zájmu o rozšíření poznání v dané oblasti (Glerup a Horst, 2014, s. 40-41). Společnost, díky demokracii, má právo vyjadřovat svůj názor a ovlivňovat směr výzkumu, vývoje a inovací. Vědci jsou optikou této racionality viděni jako někdo, kdo byl veřejností najat, aby pracoval v zájmu veřejného blaha, což není v souladu s pojetím akademické svobody.

Vědci se tomuto pojetí vědy a zodpovědnosti – sebekontrole – podle této racionality brání, a proto je nutná externí kontrola – governance – osobami a úřady, kteří vědci nejsou. Jako příklad je zde uváděna oblast rizika duálního využití (dual use), které se vztahuje na oblast UI, i když v práci Glerup a Horst (2014, s. 41) není přesně vyjmenována, vzhledem k roku publikování jejich

článku. Tato racionalita je uplatňována v souvislosti s aplikačním potenciálem UI a napojením výzkumných organizací na průmyslovou sféru.

Co je cílem snahy RRI iniciativ je probudit zájem vědců o to, formulovat své výzkumné záměry společně s různými aktéry a veřejností. Jedná se o racionalitu „**Integrační**" – jde o celý proces, ne až tak o vlastní výsupy a výsledky. Zde je vyzdvihována multidisciplinární a přímo multisektorální spolupráce. Od vědců se očekává diskuse s veřejností o nových technologiích (opět relevantní pro UI, pozn. autorky), komunikují v médiích a vytvářejí vědecké metody v součinnosti se studovanými cílovými skupinami – kooperativní přístup, který má být vzájemně obohacující, ale zároveň se očekává, že tento přístup zabrání nežádoucímu technologickému vývoji už v jeho počáteční fázi. Tady vidíme např. podporu EU v zapojení občanů v rámci tzv. citizen science.

Z výše popsaných racionalit je vidět, jak se mění přístup aktérů, kteří jsou v pozici moci (v nejširším významu to může být i veřejnost – jako někdo, na koho výsledky výzkumu přímo dopadnou, nebo z pozice toho, kdo v konečném důsledku vědu financuje) k vědcům a jaké je jejich očekávání v oblasti specifikace vědeckých problémů a způsobů jejich řešení.

Očekává se reflexivita vědců a zároveň ochota komunikovat a spolupracovat s veřejností (integrační přístup), zároveň ale, v mnoha ohledech, jde v případě UI o dodání „řešení na klíč", kdy stát, případně EU (poskytovatel grantu nebo všeobecně zadavatel) je ten, kdo vědce najímá s cílem zlepšit socio-ekonomické ukazatele dané ekonomiky (přístup užitečnosti/umožnění). Tyto přístupy koexistují – např. některé granty jsou na tzv. blue-skies (blue-sky) research, tedy neomezují, ale naopak podporují, jako granty Evropské rady pro výzkum (ERC).

Pokud se vrátíme k citátu Isidorose Karatzase na začátku kapitoly. Vše je o důvěře veřejnosti ve vědu a vědce a jejich důvěryhodnosti a odpovědnosti (accountability). V oblasti UI je potřeba spolupracovat a také veřejnost informovat, otevřeně vysvětlovat jak aplikace UI fungují a jaké jsou hrozby,

případně vyvracet mýty. Je důležité, aby probíhaly všechny tyto mechanismy. AI Act sám o sobě nezajistí vše.

3.3 Vzdělávání a odpovědnost

Vzdělání a odpovědnost jdou ruku v ruce. Jak uvádí Steen (2023, s. 56), podle morální filosofie vyplývá odpovědnost ze dvou podmínek: znalosti a agency (morálního působení), tedy stavu, intelektu, a aktivní schopnosti aplikace znalostí k ovlivnění skutečnosti a působení na okolí. Nestačí si tedy myslet, že pokud dělám jen určitou část v rámci komplexního procesu – ve vědě, výrobě, či jinde, že nejsem za tuto část odpovědný/-á. I s vědomím, že ze své pozice nemohu ovlivnit celý výsledek tohoto procesu.

Konkrétně v případě UI podle Steena (2023, s. 53) vytvářejí vědci technických a inženýrských oborů, softwaroví vývojáři a další profese nové světy a realitu, které postupně „obývají“ ostatní lidé. Z tohoto pohledu tedy nesou zodpovědnost za to, co vytvářejí, a aby do tohoto designu nezanesli svůj vlastní bias, nebo si ho aspoň byli vědomi.

Editorial časopisu Nature z června 2023 apeluje na vědce, aby budovali kulturu zodpovědné UI odspodu (bottom-up). V dubnu 2023 přijala tradiční conference v oblasti strojového učení NeurIPS (Neural Informataion Processing Systems) etický kodex,[71] který doplňuje již stávající NeurIPS kodex chování zaměřený na profesionální jednání a výzkumnou integritu.

Vydavatelské domy (viz 2.2.8), profesní organizace (viz 2.2.11) a další aktéři působící v ekosystému VaVaI, kde je vyvíjena UI, v tomto případě splňují svou povinnost “morálního agenta.” Mají znalosti a zároveň svým aktivním působením mohou ovlivnit určité chování a kulturu daného oboru. NeurIPS sestavil I etická pravidla pro hodnotitele, kteří jsou klíčovým prvkem při výběru konferenčních příspěvků.[72]

[71] K dispozici zde: https://neurips.cc/public/EthicsGuidelines.

[72] Pravidla pro hodnotitele, aby příspěvky splnili minimální stanard v oblasti etiky. K dispozici zde: https://neurips.cc/Conferences/2023/EthicsGuidelinesForReviewers.

Podobně, jako v případě etických pravidel v medicíně a biomedicíně, etická pravidla pro UI by měla zajišťovat bezpečí pro osoby, které se výzkumu využívajícího UI, účastní (jsou předmětem výzkumu osobně, nebo jejich osobní data). V tomto případě pravidla NeurIPS doporučují konzultovat tuto problematiku s vědeckými etickými komisemi na jejich výzkumných institucích, případně, pokud tyto komise nejsou ustanoveny nebo nejsou dostatečně kvalifikovány v tomto oboru, nechat si projekt posoudit panelem (interních) odborných hodnotitelů.

Tato praxe není na technických univerzitách v oboru UI zatím běžnou součástí výzkumu. Vědci při podávání grantových přihlášek či konferenčních příspěvků a odborných článků do vědeckých žurnálů provádějí sebehodnocení,[73] např. formou „zaškrtávacího" („box-ticking") dotazníku, který může být kombinován se slovním etickým sebehodnocením, které nabízí větší prostor k popisu situace, zhodnocení případných rizik a návrhu opatření. Jedná se o povinnou součást grantové přihlášky např. v programu Horizon Europe, která je též hodnocena. Evropská komise nabízí i návod, jak postupovat.[74]

Důležité je začít přemýšlet o etickém kontextu a výzkumu (celý cyklus, až po dopad/impakt, pokud to bude v daném případě relevantní) co nejdříve, ideálně hned od začátku. Příručka (EC, 2021, s. 1) doporučuje využívat metodu „Etics by design" pro aktivity, které mají vysokou inovativní přidanou hodnotu. Projekty vyvíjející či využívající systémy a technologie UI se ještě navíc musí řídit příručkou „Ethics By Design and Ethics of Use Approaches for Artificial Intelligence,"[75] kde jsou oba tyto přístupy blíže vysvětleny.

---

[73] Pokud je poskytovatelem grantu vyžadováno. Postup grantových agentur je v případě etického hodnocení projektu nejednotná – viz. prof. Veselská v 3.2

[74] Návod "Jak vyplnit etické sebehodnocení" je k dispozici zde: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf

[75] Více o přístupu k hodnocení etického dopadu ve 3.4.

Praxe je zatím spíš taková, že se pár minut před odevzdáním teprve zaškrtávají políčka formuláře, ale už není čas na řádné zakomponování etických prvků a promyšlení souvislostí do textu. Na tom potom vědci zbytečně ztrácejí „body" při hodnocení odborným panelem.[76] Zkušenosti z praxe sdílela na konferenci pořádané Masarykovou univerzitou v rámci „Grants week" 22. 11. 2023 (MUNI, 2023) prof. Renata Veselská, členka poradní Evropské skupiny pro etiku ve vědě a nových technologiích (European Group on Ethics in Science and New Technologies, EGE), předsedkyně Etické komise pro výzkum MUNI, a také jedna z řešitelů projektu TA ČR: INSURE – Interdisciplinární podpora etiky ve výzkumu. V rámci tohoto projektu vyšly publikace: Posuzovací guidelines pro etické komise a Průvodce vědeckou integritou. Projekt sice není specificky zaměřen na umělou inteligenci, ale publikace jsou v mnoha ohledech univerzálně použitelné.

3.4 Metody etického sebehodnocení – od formulářů ke spolupráci s cílovými skupinami

Kombinace obou přístupů, „zaškrtávacího způsobu" s rozsáhlejším textem tak, jak je popsáno výše, je pravděpodobně ideální. Správně sestavený formulář pro etické sebehodnocení/evaluaci, který poskytuje grantový poskytovatel nebo je možné si udělat interně vlastní pro lepší kontrolu a srovnání, nabízí přehled všech kroků, které je potřeba v té které fázi přípravy projektového záměru/výzkumu udělat a promyslet. Případně domluvit si externí konzultaci – s vědeckou etickou komisí[77], interním peer review panelem, kolegy, zajistit si

[76] Zkušenosti z vlastní praxe v hodotitelských komisích projektů Horizon Europe agentury REA (European Research Executive Agency– v roli „Mladého pozorovatele", v rámci stáže Young observers v agentuře REA, kam jsem byla v roce 2022 vybrána. Stáž probíhala od 11/2022 – 02/2023.
Seznam účastníků zde: https://research-and-innovation.ec.europa.eu/document/download/ac06dbaa-11fe-4394-991f-08b7392058f1_en?filename=Young%20Observers%20list%202022%20v3.pdf

[77] V projektech UI není tento postup běžný. Jedná se spíše o možnost do budoucna založenou na dobrovolné iniciativě vědců, pokud bude komise podobné poradenství v žádoucím rozsahu a expertize schopna a ochotna poskytnout.

poradenství externě, nebo přímo zapojením SSH vědců do projektového konsorcia a tím i do přípravy projektu, což je v Horizon Europe projektech již standardním požadavkem.

Metoda ethics by design, popsaná v příručce (EC, 2021, s. 11) si klade za cíl předcházet možným rizikům již ve fázi přípravy díky proaktivnímu využití všech základních principů (respekt k lidskému jednání; soukromí; ochrana osobních údajů a jejich správa; spravedlnost; individuální, sociální a environmentální blaho; transparentnost; odpovědnost a dohled) jako určitých filtrů, kterými se na plánovaný UI systém (a celý vývojový proces) díváme. Pokud chystané UI řešení porušuje byť jen jeden ze základních principů, bude považován za neetický.

Příručka upozorňuje na to, že **ne vše, co je možné udělat, by se mělo udělat**. Etické požadavky ohledně transparentnosti, zodpovědnosti a lidského dohledu nejsou cílem, ale mají se vztahovat přímo k **architektuře technického designu**. Dále doporučuje zamýšlet se, jak by bylo možné dané řešení záměrně nebo i náhodou zneužít a vždy využívat nástroje, které zajistí splnění všech etických principů a nakládat s osobními daty v souladu s GDPR (na to je samostatná příručka Note on Ethics and Data Protection). Příručka má v závěrečné části seznam všech kroků, které je potřeba podniknout k úspěšnému naplnění daných etických principů.

OECD[78] také doporučuje tzv. „**by-design**" přístupy (např. „Ethics-by-design" nebo „Sustainability-by-design"). Sebe-regulaci vycházející z praxe určitého oboru nebo průmyslu nazývá „upstream governance" – tedy opět směr zdola nahoru. To může pozitivně ovlivnit celé dodavatelsko-odběratelské řetězce tam, kde se daný obor vyvíjí příliš rychle a není zatím regulován shora (top-down). Vlády mohou relativně rychle podpořit pozitivní směr vývoje nových technologií vyhlášením tendrů a veřejných soutěží – např. jako Bílý dům tzv. na

[78] OECD Science, Technology and Innovation Outlook 2023 (2023b).

technologie potvrzující demokracii (democracy affirming technologies) v roce 2021 nebo technologie podporující právo na soukromí (privacy) (WH, 2022).

Jako další metodu nabízí také **co-design**, při kterém se využívají participativní mechanismy, aby se cílové skupiny – obyvatelé nebo malé a střední firmy (SMEs) - mohli podílet na designu daného technologického/technického řešení. Aby byl tento systém transparentní ve vztahu k veřejnosti, doporučuje OECD vytvořit procesy, které budou obsahovat jasně stanovené principy, standardy, metodiky a kodexy (tzv. **soft-law regulace**) a zajistit jejich standardizaci.

Posledním navrhovaným přístupem je tzv. **compliance** – tedy shoda, dodržování pravidel a standardů. „Vytvoření mechanismů dohledu nad implementací a dodržováním standardů, vč. auditů třetích stran, zda je dodržována technologická „governance" jako součást infrastruktury efektivní kontroly kvality (OECD, 2023c)." Do tohoto ekosystému shody patří další mechanismy (např. v oblasti odpovědnosti za škody) a aktéři jako externí etické komise, pojišťovny, vládní agentury, atd.

Součástí přípravy a procesu sebehodnocení etiky a impaktu výzkumu (projektového záměru) a nejjednodušší metodou, je naučit se pokládat si (i nepříjemné, náročné) **otázky** a být otevřeni možnostem a komentářům kolegů, ideálně s jinou oborovou specializací.

Otázky na nejvyšší úrovni, které berou v potaz i budoucí technologický vývoj a dopad UI systémů na lidskou společnost, které byly v roce 2017 definovány Future of Life Institute (FLI, 2024) v době, kdy svolal konferenci předních myslitelů a vědců v oblasti UI do Asilomaru (tzv. Asilomarské principy vývoje UI - viz 3.2):

„Jak můžeme zajistit vysokou míru robustnosti u budoucích systémů UI, aby dělaly přesně to, co zamýšlíme, bez poruchovosti nebo rizika, že podlehnou kybernetickému útoku?

Jak můžeme zajistit růst prosperity díky automatizaci, při zachování lidských zdrojů a smyslu života?

Jak můžeme zajistit aktualizaci legislativních systémů, aby byly férové a účinné a schopné držet krok s vývojem UI, a zároveň řídit rizika spojená s UI?

S jakými morálními hodnotami by měla být UI sladěna (aligned) a jaký právní a etický status by měly mít?"

Dalším příkladem je sada otázek, na které doporučuje si odpovědět Ada Lovelace Institute ALI (2020, 29. dubna) v případě procesu hodnocení algoritmického systému:

- Jaká jsou zdrojová data systému?
- Kdo ho řídí?
- Jaké jsou možné podmínky pokud bude algoritmus nasazen i pro úlohy, na které nebyl určen nebo v jiném kotextu než na který byl vytvořen a testován?
- Proč je systém vyvíjen?
- Jaké byly možné alternativy a proč bylo vybráno zrovna současné technické řešení?
- Jsou nějaké specifické skupiny, které budou daným řešením ovlivněny/postiženy?
- Podíleli se na technickém řešení soukromí investoři?
- Jaká je logika systému?

V případě odpovědného přístupu v oblasti robotiky (doslova roboti pro dobro – „Robots for good"), navrhují autoři Šabanović, S., Charisi, V. & Belpaeme et al. (2023), aby se vývojáři zamysleli nad následujícími dotazy.[79] Roboti nejsou

[79] Dotazy jsou oproti původnímu zdroji, kde jsou velmi podrobné, autorkou zkráceny. Překlad vlastní.

inherentně dobří či špatní, proto mnoho otázek směřuje přímo na záměry a cíle vývojářů. Lze z toho usoudit, jak nad svým produkterm přemýšleli, jestli je tam riziko biasu, jestli jsou některé oblasti, které nevyřešili nebo se nad nimi nezamysleli atd.:

- Jak jste dospěli k názoru, že vaše pojetí „dobra“ povede k pozitivním změnám ve společnosti? Kterou z cílových skupin je vaše řešení inspirováno/motivováno? Je v souladu s cílovou skupinou nebo hranicí vědeckého poznání (tedy vědeckými cíli) či jinými motivy?
- Pro koho je vaše řešení „dobré“? Specifikujte, jak vypadá váš konečný uživatel?
- Pro koho naopak není vaše řešení dobré?
- Máte definovanou konkrétní matici či indikátory jak měřit „dobro“ vašeho řešení?
- Kteří partneři se spolupodíleli na vývoji vašeho řešení?
- Jaká omezení váš robot má, v tom, co je schopen dělat.
- Jaký je nejhorší scénář, co nejhoršího se může stát? Jaké je největší možné riziko pro uživatele vašeho řešení?
- Je nezbytné zavést nějaké regulace či jiná opatření v oblasti ochrany spotřebitele v zájmu uživatelů vašeho řešení?
- Jaké jsou ochranné prvky vašeho robota, aby byl bezpečný pro uživatele?
- Jak jste zajistili transparentnost vašeho řešení (zabránili zkreslení).

3.5 Etika jako tacitní znalost inženýrů a developerů?

Od etiky UI přes široce definované principy důvěryhodné UI (Trustworthy AI) až po konkrétnější způsoby, jak přistupovat k přemýšlení a posuzování impaktu vlastního výzkumu a vývoje. Jedná se v podstatě o vzdělávání, senzitizaci a rozvoj schopností předvídat, co se může stát. Předjímat možná rizika a negativa, stejně jako pozitiva vlastního výzkumu. Ústřední otázka by měla znít,

proč daný výzkum dělám (MUNI, 2024)?[80] Jaké metody zvolit, aby si vědci, případně ideálně ještě studenti inženýrského (magisterského) studia, mohli vyzkoušet, jak mohou o dopadu svého výzkumu přemýšlet (studenti např. o své diplomové nebo doktorské práci, aby to bylo konkrétnější), jaké otázky si pokládat a kde a případně s kým na ně hledat odpovědi?

Nakonec díky tomu, že by si zažili tento „brainstormingový" kreativní proces, by se jim to dostalo lépe pod kůži. Cílem je budovat **tacitní dovednosti**[81] v oblasti etiky UI a vědecké integrity a dovednosti inter/multidisciplinární komunikace, aby se postupně staly naprosto přirozenou součástí jejich vědecké práce nebo zaměstnání mimo akademickou sféru. I když z rozhovorů vyplynulo, že některý výzkum je čistě teoretický, mnoho výzkumných projektů je interdisciplinárních a v praxi je potřeba naučit se komunikovat (rozvíjet tuto soft-skills dovednost) a pochopit/respektovat vědce z ostatních oborů (viz podkapitola 1.1.5).

Představivost je také důležité rozvíjet. Hodí se nejen při popisu impaktu a možných rizik, zvláště do grantových přihlášek ERC nebo jiného blue-skies výzkumu. Představivost, která se opírá o reálné zkušenosti a znalosti (teoretické, praktické, tacitní) a kvalifikovaný odhad možných rizik.

---

[80] **Modelové situace** – proč někteří vědci výzkum dělají (tedy, jak by to být nemělo), jsou ve výstupech projektu INSURE Jedná se sice o projekty v jiné oboru než UI, ale analogie je využitelná i zde.
Např. „Rozhodl jsem se zkoumat téma mentální reprezentace smrti u dospívajících, protože je o tom málo napsáno a cítím tu určitou šanci na publikaci. Vytvořil jsem si tedy schéma rozhovoru pro kvalitativní výzkum a plánuji s ním oslovit děti ve volnočasovém zařízení, kde pracuji. Budu se jich ptát na jednoduché věci – jestli někdy přemýšlely o tom, že by zemřel někdo z rodičů, jestli jsou naživu všichni prarodiče a tak. Naši etickou komisi s tím vůbec nebudu obtěžovat, vždyť o nic nejde." Servisní středisko pro e-learning na MU, Informační systém Masarykovy univerzity. (n.d.). INSURE: Průvodce vědeckou integritou, Lékařská fakulta MU. Masarykova Univerzita. https://is.muni.cz/do/rect/el/estud/lf/js23/vedecka_integrita/web/pages/trasa_8.html

[81] Koncept "tacitní dovednost" jako termín vymyslel maďarský filosof Michael Polanyi. Jedná se o popis pochopení podstaty toho, jak něco funguje, JAK se vykonává – procesu, než znalosti (teoretické - CO). Je možné se tacitním dovednostem naučit, osvojit si je tak dokonale, že je vykonáváme spíš podprahově, bez přemýšlení, téměř automaticky. Zároveň je často ani nevnímáme jako konkrétní dovednost, kterou bychom byli schopni popsat slovy, definovat, např. na životopise (tento stav se nazývá Polanyiho paradox).

Nabízí se možnost klást si vzájemně otázky – viz výše. Ještě lepší možností by bylo například pokusit se zorganizovat mock peer review se zástupci a zástupkyněmi různých oborů, genderu a menšin. To by mohl být návrh např. pro doktorandy určité výzkumné skupiny nebo v rámci katedry, atd.

Zatím je výuka společenských věd na technických unierzitách spíše teoretická, filosofická nebo etiku UI univerzity řeší jako volitelný online kurz pro studenty z nabídky prg.ai nebo přímo anglickou verzi: https://ethics-of-ai.mooc.fi/ (PRGAI, 2024).

Na základě konzultace s B. Pelcovou z prg.ai jsem zjistila, že tento kurz využívají následující VŠ: Univerzita Karlova, Slezská univerzita v Opavě, UJEP a Univerzita Palackého a Mezinárodní institut IT a programování. Je to skvělý základ, na který by mohly navazovat praktické workshopy, ideálně mezi studenty technických a SSH oborů. V rámci RAI (Responsible AI) neformální pracovní skupiny na Katedře počítačů podobnou spolupráci zvažujeme, na úrovni doktorského studia, tedy v malé skupině.

Kromě kombinace výše popsaných technik, tedy:

- Teoretického základu Etiky UI v podobě online kurzu;
- Brainstormingu relevantních otázek vztahujících se k vlastnímu výzkumu nebo diplomové práci a
- Mock peer-review (s výzvou sestavit daný panel co nejpestřejší).

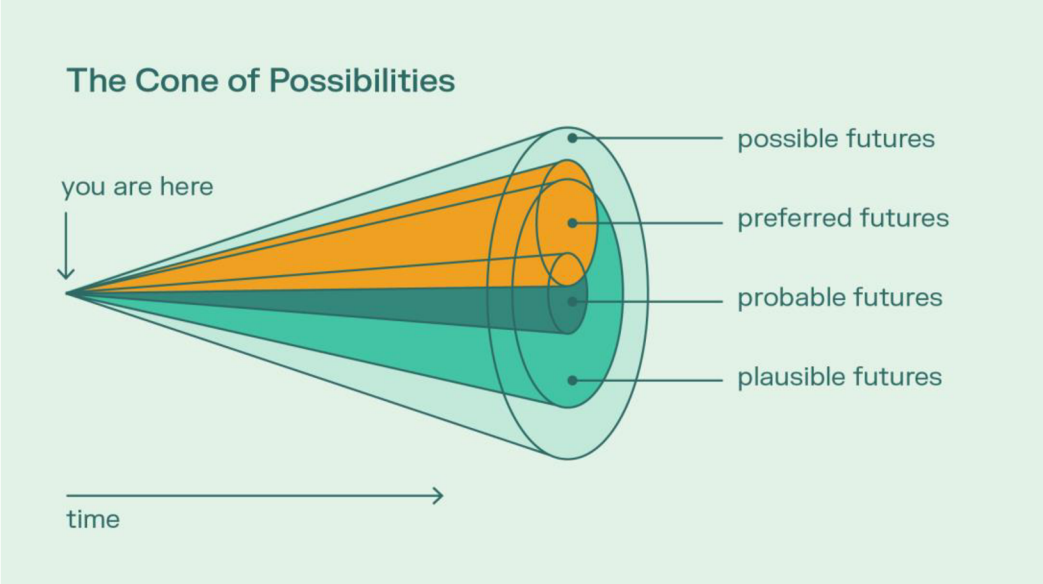
Je možné využít ještě dalších technik „předjímání budoucnosti“, které popisuje v knize Etika budoucnosti Bowles (2021, s. 37 - 43): „Běžným modelem futurologie je kužel budoucnosti, který používá analogii světla vyzařovaného pochodní (nebo spíše baterkou)“.[82]

Světlo je potenciální budoucnost. Čím dál na časové ose se nachází, tím méně ostřejší námi preferovaná budoucnost je. Předvídatelnost se snižuje.

[82] “Původně vymyšlený vojenským stratagem Charlesem Taylorem a od té doby upravený několika futuristy včetně Josepha Vorose (Bowles, 2021, s. 37).”

V jasném středu je pravděpodobná budoucnost, kolem ní ve větším okruhu je stále přijatelná budoucnost, nejširší je potenciální budoucnost.

Budoucnost, kterou si skutečně přejeme je mimo hlavní střed. Jedná se o velmi zjednodušený model, který ale umožňuje představit si, že tím, co dnes dělám, budoucnost aktivně ovlivňuji nebo aspoň tu část, kterou mohu. Kužel budoucnosti tedy „může být užitečnou nápovědou pro práci na strategii i oporou morální představivosti, osvětluje totiž různé budoucí trajektorie a tímu umožňuje vybrat si preferovanou budoucnost, o kterou bude usilovat. (Bowles, 2021, s. 38).



Obrázek 12 Kužel budoucnosti (obrázek z www.delve.com)

Další technikou, kterou Bowles navrhuje (2021, s. 39) je namalovat si „kolo budoucnosti" Jeromeho C. Glenna, které nabízí „**strukturovaný brainstorming**. Začneme u původního trendu, například u zvolené technologie, a pak v kruhu kolem něj popisujeme některé jeho možné důsledky.“ Výsledek vypadá jako jednoduchá květina, kdy okvětní lístky přímo napojené na střed představují pravděpodobnou budoucnost. Na každý z nich potom v další vrstvě přidáváme potenciální důsledky druhého řádu a vytvoříme tak druhý kruh. Lze navázat ještě další vrstvou, např. úplně hypotetického scénáře (s. 39). Jednotlivé scénáře

propojujeme, tvoříme uzly podle toho, jak spolu souvisejí. „Evokování společných vizí budoucnosti je základním kamenem morální představivosti a zásadním způsobem, jak na nezamýšlené důsledky upozornit (s. 40). (A třeba si lze pomocí této techniky zkusit vymyslet i varianty řešení tramvajového dilematu).

Dalším stupněm je osvojení si metody Ethics by design (viz 3.4), který navrhuje Evropská komise ve své příručce (EC, 2021, s. 11) a použít ji při psaní projektových návrhů ve výzvách Horizon Europe (nebo popsat rizika reálného vlastního výzkumu, disertační či diplomové práce).

Co mají výše popsané návrhy společné je, že se jedná o zcela dobrovolnou aktivitu. Prof. Veselská v rozhovoru (Marušáková, 2023) potvrdila, že preferuje, aby byly její přednášky z bioetiky nepovinné. Přesto, nebo právě proto, jsou extrémně populární napříč celým spektrem oborů na MUNI. Popsala, jak organizace výuky probíhá: ...

„...Semestr končíme kolokviem a po studentech chci aktivní a tvůrčí přístup. Na kolokvia přichází každý z nich s konkrétním etickým dilematem, které osobně považuje za důležité, a spolu s ostatními je diskutujeme. Tohle je pro mne ta nejzajímavější a nejpřínosnější část, kdy mohu sledovat uvažování studentů a jejich etickou reflexi toho, s čím se v životě setkávají. Dává mi to zpětnou vazbu, která se samozřejmě promítá do dalšího obsahu obou předmětů.[83] *A také radost z vědomí, že si studenti pořád uchovávají schopnost kritického myšlení, i naději, že jejich morální kompasy fungují.“*

Změny v oblasti výuky se chystají i na FEL ČVUT, kde je ve fázi přípravy k akreditaci nový studijní magisterský program, jehož cílem bude poskytovat vzdělání v hlavních disciplínách umělé inteligence s ohledem na jejich

[83] Jedná se o předměty Bioetika s podtitulem „Etika života“ a v jarním semestru s podtitulem „Možnosti na hraně“.

společenský dopad. Do povinného kurikula budou zařazny i společenské vědy, což reflektuje skutečnost, že UI prostupuje do společnosti a je potřeba studenty senzitivizovat s ohledem na možná rizika a související etické otázky dopadu UI na lidská práva. „Tento prostup dále otvírá otázky sociálně-technické robustnosti modelů strojového učení, jež jsou klíčové pro jejich bezpečené nasazování v dynamických a komplexních společenských prostředích s možnou přítomností protivníků (z interního dokumentu ČVUT, ukázka použita se souhlasem garanta studijního programu).

3.6 RRI na univerzitách – jak na změnu interních procesů a kultury?

Výzkumné organizace se postupně mění a reagují na společenskou poptávku. Jednak mají určitou společenskou odpovědnost, která vyplývá ze Zákona o VŠ č. 111/1998 (viz 2.1.2), jednak musí být obory relevantní, aby jejich absolventi obstáli na pracovním trhu a měli všechny nezbytné dovednosti. Zároveň musí být i dostatečně atraktivní pro zahraniční i české studenty a studentky a společenský rozměr technických oborů by v tomto směru motivující skutečně mohl.

Jako výzkumné instituce a veřejné vysoké školy musí zajistit transparentnost interních procesů (i v rámci např. HR Award) a určité nezbytné agendy, včetně podpory RRI. Předpokladem je deklarace etických hodnot v etickém kodexu, ustanovení etických komisí v rámci fakult i celouniverzitní vědecké etické komise. Jak ještě podpořit změnu prostředí, kultury a podpory praktického přístupu k integraci etických principů do výuky a výzkumu jednotlivých oborů v rámci dané instituce?

Vytvoření interních pravidel a postupů v oblasti RRI, které budou v souladu s principy etického vývoje např. UI, vzdělávání a zlepšení komunikace s vědeckou etickou komisí, jejich zavedení do praxe a akceptace vědeckou komunitou nebude otázka několika měsíců, ale spíš let a poté kontinuální proces

zlepšování a „údržby“. Tímto směrem se například ubírá Technická univerzita ve Vídni. Dr. Marjo Rauhala, informovala na konferenci EARMA 2023 v Praze, jak postupuje s cílem podporovat znalosti a dovednosti vědců technických oborů (STEM), aby byli schopni nahlížet na svou práci z etického hlediska a komunikovat o ní s odborníky z jiných vědeckých disciplín.

Na druhou stranu je to způsob, jak zároveň vzdělávat i členy etické komise TU Wien. Obě strany se tak učí a rozvíjejí dovednosti, které jsou nezbytné k zajištění špičkové úrovně výzkumu a schopnosti určité sebe-regulace a sebe-reflexe. Důležité je zdůraznit, že se nejedná o externí regulaci od zřizovatele, státu či rektorátu (top-down), ale jejich zapojení je zcela dobrovolné. Mohou konzultovat své projektové záměry i na začátku přípravné fáze a připravit tak lepší projekt. Podobný postup, jak popsala prof. Veselská (2023, 22. 11.), funguje na MUNI.

Jedná se tedy o přístup k prosazování změn „odspodu“, s tím, že celý projekt/proces zaštiťuje „shora“ rektorát univerzity tak, aby se vzdělávali nejen vědci, ale i odborníci v etické komisi, aby se od sebe učili navzájem a také se pochopili a respektovali. Vzniká zde jedinečná organizační kultura.

Nové technologie se vyvíjejí tak rychlým způsobem, že je složité pro etické komise a členy různých hodnotících výborů tyto změny sledovat a vůbec porozumět jejich možným souvislostem a případným hrozbám či inovačním příležitostem. Evropská komise proto podporuje vzdělávací projekty pro vědce se specializací v netechnických oborech nebo STEM vědce ve vzdělání např. v aplikované etice, jako projekt iRECS.[84]

Podpůrnou metodikou, kterou doporučuje i Karatzas, je příručka RRI ALLEA – Evropský kodex integrity výzkumu. V roce 2023 vyšlo revidované

[84] Improving Research Ethics Expertise and Competences to Ensure Reliability and Trust in Science, project Horizon Europe https://cordis.europa.eu/project/id/101058587

vydání. Je možné také využít nástroje vyvinuté v rámci Horizon2020 – např. praktický **RRI toolkit: https://rri-tools.eu/**.

Partneři ze sféry výzkumných organizací, ale i poskytovatelů výzkumných grantů, konsorcia čtyřletého projektu SOPs4RI (Standard Operating Procedures for Research Integrity), financovaného EU v rámci Horizon 2020, vytvořili set nástrojů – toolbox – které mohou pomoci v nastartování a definici interních procesů v rámci výzkumné integrity. Výzkumným organizacím je určeno přes 120 nástrojů v rámci devíti témat, poskytovatelům nabízejí 29 nástrojů v rámci 6 témat - **https://sops4ri.eu/toolbox/**.

Dalším, v červnu 2023 ukončeným, projektem Horizon 2020 byla ETHNA **https://ethnasystem.eu/results/**. Cílem tohoto projektu bylo implementovat a zajistit fungování interních manažerských procesů a systémů zodpovědného výzkumu a vědecké integrity (RRI) v šesti Evropských centrech pro vyšší vzdělání, financování a výzkum (HEFRC – European Higher Education, Funding and Research Centres).[85]

Další relevantní zdroje a odkazy: Kretser, A., Murphy, D., Bertuzzi, S. et al. (2019) napsali skvělý článek k principům vědecké integrity, který obsahuje i příklady dobré praxe. Principy jsou velmi dobře popsány a vysvětleny. A velmi užitečnou příručkou k tématu vědecké a výzkumné integrity a etiky je kniha Fostering Integrity in Research od NIH – National Library of Medicine. Platforma pro podporu RRI a etiky, je Embassy of good science: https://embassy.science/wiki/Main_Page.

Etické komise mohou fungovat na bázi poradních orgánů ve smyslu dobrovolných konzultací např. k projektovému záměru nebo k jeho určité fázi, pokud je jejich složení pro posouzení daného záměru vhodné a odpovídá potřebné šíři expertizi. Je to prevence, ale zároveň i možnost získat feedback od

[85] V rámci tohoto projektu vzniklo mnoho výstupů a publikací – např. González-Esteban, E., Feenstra, R. A., & Camarinha-Matos, L. M. (2023). Ethics and responsible research and innovation in practice: The ETHNA System Project. Springer Nature.

vědců s jinou vědeckou, profesní a životní zkušeností. Tedy svůj výzkumný záměr na základě toho vylepšit, lépe popsat nebo naopak již neřešit „slepé uličky" a zajistit vyšší efektivitu při vlastní realizaci, ale i vyšší šance grant získat. Je to oblast, která může hodnocené projekty v celkové vysoké konkurenci jinak též vědecky excelentních projektů vyzdvihnout a je to přesně ta forma podpory, kterou zmiňoval na začátku kapitoly 3 Karatzas, která některé výzkumné organizace kvalitativně odlišuje od jiných.

Na druhou stranu může být pro některé výzkumné skupiny složité zajistit si konzultace s vědci z jiných oborů, pokud nebude možná konzultace s etickou komisí, a případně je i zaplatit, pokud se s nimi třeba přímo nepočítá při budoucí realizaci projektu.

U profesionální vědecké komunity je důležité nastavit interní kulturu výzkumné organizace a management rizik tak (protokol, formální postupy, jak v dané situaci reagovat, kdo je zodpovědný k potvrzení a prověření dané situace atd.), aby vědci a zaměstnanci neměli obavu vyjádřit se nebo nebyli perzekuovaní za své názory, pokud zjistí nějaké malfunkce, bias nebo další negativní aspekty vyvíjených UI produktů či jiné eticky sporné situace.

Příkladem vědkyň a vědců, kteří však za své názory, které nebyly většinovou komunitou přijaty, a dokonce někteří ztratily své místo, aby se nakonec ukázalo, že jejich obavy byly oprávněné, jsou Joy Buolamwini , Timnit Gebru nebo Josepha Weizenbaum, který v roce 1966 vytvořil první chatbot Eliza na Massachusettském technologickém institutu (MIT).

Neřešení interních problémů a nesouladu mezi morálním přesvědčením odborníků a managementu těchto organizací může vést v krajním případě k tzv. whistleblowingu, tedy upozornění příslušných kontrolních orgánů zaměstnanci na interní postupy a praktiky dané organizace, případně zveřejnění důležitých interních dokumentů i za cenu možné vlastní persekuce.

Podle prof. Veselské (Marušáková, 2023) je: ...

„... osobní mravní integrita jeden celek, a pokud člověk nemá nastavený vnitřní morální systém, pokud se například neeticky chová k účastníkům výzkumu nebo ke studentům při výuce, tak je to obvykle spojeno s tím, že nedodržuje morální principy ani v ostatních rovinách své práce.“

Mnozí se naopak zapojují do nejrůznějších národních, mezinárodních, a především interdisciplinárních projektů, inciativ a politik, aby pozitivně ovlivnili další vývoj a směřování výzkumu umělé inteligence a způsobu, jak je aplikována v jednotlivých oblastech, které mají přímý vliv na člověka, ale i na společnost a složité vztahy a procesy v ní. „Umělá inteligence musí být vybavena lepším porozuměním lidské inteligenci, hodnotám a potřebám, aby sloužila nejlepším zájmům lidstva (CHIA, 203).“

Joy Buolamwini založila The Algorthmic Justice League v MIT Media Lab. Upozornila na rasovou diskriminaci softwaru pro rozpoznávání obličejů, který nebyl schopný identifikovat tvář jiné než světlé pleti.

Timnit Gebru, bývalá vývojářka firmy Google, byla propuštěna poté, co upozorňovala na nedostatky a nedostatečné zajištění etického přístupu k vyvíjeným produktům využívajícím UI např. v oblasti biasu. Technologie na rozpoznávání obličejů, kterou Google vyvinul diskriminoval lidi tmavé pleti, zejména však Afroameričanky. Data, na kterých byl systém natrénován (machine learning) neobsahovala dostatečný počet fotografií žen z minoritních etnických skupin. 12 měsíců po svém propuštění založila nezávislý výzkumný institut pro studium etiky AI velkých technologických korporací (Perrigo, 2022, January 18).

ZÁVĚR

Cílem práce bylo zodpovědět výzkumnou otázku: Jaké kontrolní mechanismy ovlivňují výzkum a výstupy v projektech využívajících nebo vyvíjejících umělou inteligenci, aby byly v souladu s etickými principy?
Vycházela jsem z následujících hypotéz:

Hypotéza 1: Propojení oborů UI a etiky zatím zcela chybí, protože se nepropisuje do vědecké praxe ani managementu projektů, kde je UI předmětem výzkumu nebo je využívána v kontextu jiných oborů a dopadů v celé řadě oblastí a profesí.

Zjištění 1: Vývoj v oblasti etiky UI a vědecké integrity (RRI) se velmi rychle vyvíjí v globálním měřítku, ale do místní kultury na vědeckých pracovištích se tyto změny propisují pomaleji. Práce vychází z teorie v oblasti etiky UI a jejího postupného přenosu do reálného vědeckého prostředí – v rámci empirického výzkumu (rozhovory) a pozdější aplikací v praxi, při přípravě projektových žádostí a zpracování etických sebehodnocení výzkumných záměrů.

O etické otázky vědci zájem mají a nečekaným pozitivním výstupem tohoto výzkumu bylo založení neformální skupiny Responsible AI, která diskutuje témata spojená s etikou UI v rámci pravidelných schůzek.

Hypotéza 2: Vývoj UI, která bude společensky prospěšná a v souladu s etickými principy je možné zajistit interdisciplinání spoluprací s filosofy a etiky v rámci spolupráce na společných projektech?

Zjištění 2: Přístup k řešení spolupráce v oblasti etiky a impaktu výzkumných projektů v oblasti UI ve formě spolupráce partnerů z různých vědeckých disciplín, tedy že etik, filosof/SSH bude přítomen v přípravné fázi nebo v rámci vědecké rady projektu či projektového balíčku není tak efektivní. Tento transfer znalostí a interdisciplinarita (vzájemné intelektuální obohacení) se nedějí automaticky a v takové míře, jak jsem předpokládala.

Je potřeba tuto spolupráci a komunikaci v rámci projektů aktivně podporovat, ideální ale je, pokud vědec či přímo hlavní řešitel sám nastuduje problematiku druhého oboru nebo aspoň jeho základy. To je podle J. Kulveita „skutečná" interdisciplinarita a lze ji aplikovat i v oblasti etiky UI v realitě inženýrských, technických vědců. Zajistí se tak lepší komunikace, pochopení odborné terminologie a vyšší kvalita vzájemného porozumění s experty na danou problematiku, respekt a celkově lepší a efektivnější spolupráce.

Hypotéza 3: Lze nalézt souvislost mezi kontrolními mechanismy a zlepšením kvality vyvíjených systémů UI v oblasti etiky a pozitivního impaktu na společnost?

Zjištění 3: Ano, mechanismy jako financování a hodnocení vědy (a mnoho dalších) ovlivňují chování a přístup vědců i k tématu etiky UI, pokud se o ni nezajímají sami, ze své přirozené touhy a v souladu s principy RRI. Jedná se potom spíše o top-down přístup, externí tlak. Také zde platí souvislost se „skutečnou" interdisciplinaritou. Pokud mají vědci sami odborný vhled do etického rozměru a společenského kontextu svého výzkumu, mají vyšší vnitřní motivaci se těmito otázkami zabývat. Ideální stav by byl, kdyby se toto „etické uvažování" postupně stalo tacitní znalostí vědců v oblasti UI. Negativem ale je, pokud nedojde k vnitřnímu ztotožnění vědců s touto problematikou, budou etiku považovat za nutné zlo, překážku, omezení, místo něčeho pozitivního, co jim umožní přemýšlet jinak nad svým výzkumem a oborem a zlepšit tak i kvalitu vlastních vědeckých výstupů.

V průběhu výzkumu jsem měla možnost poznat klíčové aktéry a mechanismy, kteří v rámci tohoto výzkumného prostředí spolu interagují a ovlivňují práci vědců i manažerů a administrátorů (RMAs). V DP je tedy jednotlivě popisuji. Aktéry jsou globální, mezinárodní instituce, Evropská unie, Rada Evopy, členství v dalších např. profesních organizacích, univezitní aliance, Ministerstva (státní správa). Mechanismy jsou financování, hodnocení, legislativa, vzdělávání, atd.

Praktickým výstupem je potom lepší schopnost autorky zapojit se do přípravné fáze vědeckých projektů (tzv. pre-award) v oblasti evaluace etických otázek spojených s výzkumem a vývojem UI, impaktu a prevence rizik. A lepší navigace komplexním prostředím VaVaI UI.

Seznam literatury a zdrojů

Ada Lovelace Institute, the, ALI (2020, 29. dubna). *Examining the Black Box. Tools for assessing algorithmic systems. Identifying common language for algorithm audits and impact assessments.* Citováno 9. ledna 2024. Dostupné z: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Ada Lovelace Institute, the, ALI (2020, 15. října). *Transparency mechanisms for UK public-sector algorithmic decision-making systems. Explainer for Government, local government, policymakers and researchers. A review of existing UK mechanisms for transparency, and their contribution to making public information relating to the implementation of algorithmic.* Citováno 9. ledna 2024. Dostupné z: https://www.adalovelaceinstitute.org/wp-content/uploads/2020/10/Transparency-mechanisms-explainer-1.pdf

Alignment of Complex Systems Research Group (ACS), Charles University in Prague (n.d.). Citováno 23. února 2024. Dostupné z: https://acsresearch.org/about

ALLEA (2023) *Evropský kodex integrity výzkumu – revidované vydání 2023 (čeština).* Berlín. DOI 10.26356/ECOC-Czech

Allen, D., Hubbard, S., Lim, W., Stanger, A., Wagman, S., & Zalesne, K. (January 2024). *A Roadmap for Governing AI: Technology Governance and Power Sharing Liberalism*. Ash Center for Democratic Governance and Innovation at Harvard University. Citováno 12. února 2024. Dostupné z: https://ash.harvard.edu/publications/roadmap-governing-ai-technology-governance-and-power-sharing-liberalism

Amnesty International. (2021, October 11). *'The Great Hack': Cambridge Analytica is just the tip of the iceberg*. Citováno 13. března 2024. Dostupné z: https://www.amnesty.org/en/latest/news/2019/07/the-great-hack-facebook-cambridge-analytic

Artificial Intelligence Index, Stanford Institute for Human-Centered Artificial Intelligence (2023). *Artificial Intelligence Index Report 2023.*

Association for the Advancement of Artificial Intelligence (2024) *AAAI Ethics and Diversity*. Citováno 1. února 2024. Dostupné z: https://aaai.org/about-aaai/ethics-and-diversity/

Association for Computing Machinery (2023). *ACM FAccT Conference 2024*. Citováno 19. března 2024. Dostupné z: https://facctconference.org/2024/

Barnes, B., Koblin, J., & Sperling, N. (2023, July 14). *Hollywood actors join writers on strike, bringing industry to a standstill*. The New York Times. Citováno 21. března 2024. Dostupné z: https://www.nytimes.com/2023/07/13/business/media/sag-aftra-writers-strike.html

Barták, O. Deeply (blog). (2024) *Co je to generativní umělá inteligence? Vše co musíte vědět.* (n.d.). Citováno 12. ledna 2024. Dostupné z: https://deeply.cz/blog/co-je-to-generativni-umela-inteligence

Bartneck, C., Lütge, C., Wagner, A. R., & Welsh, S. (2021). *An introduction to ethics in robotics and AI*. Dostupné z: SpringerBriefs in ethics. https://doi.org/10.1007/978-3-030-51110-4

Bartoš, I. (2023). Vládní Program Digitalizace 2018+ Digitální ekonomika a Společnost. Citováno 11. ledna 2024. Dostupné z: https://vlada.gov.cz/assets/ppov/rvis/zapisy_rvis/DES-2023.pdf

Bezpečnostní informační služba (BIS) (2023). *Výroční zpráva 2022.* Citováno 12. ledna 2024. Dostupné z: https://www.bis.cz/vyrocni-zpravy/vyrocni-zprava-bezpecnostni-informacni-sluzby-za-rok-2022-2cd547c8.html

Boddington, P. (2020). Normative Modes: Codes and Standards. In Dubber, M. D., Pasquale, F., & Das, S. (Eds.). *The Oxford Handbook of Ethics of AI* (s. 125-140). New York: Oxford University Press.

Bowles, C. (2021). *Etika budoucnosti.* Praha: Academia

Brauner, J., & Chan, A. (2023, August 10). *AI poses Doomsday Risks—But that doesn't mean we shouldn't talk about present harms too*. Time.com, citováno 28. února 2024. Dostupné z: https://time.com/6303127/ai-future-danger-present-harms/

British Telecom, BT Group (2024) *Applying responsible tech principles across the value chain.* Citováno 14. března 2024. Dostupné z: https://www.bt.com/about/digital-impact-and-sustainability/championing-human-rights#value-chain

Budd, J., Miller, B. S., Manning, E., Lampos, V., Zhuang, M., Edelstein, M., Rees, G., Emery, V. C., Stevens, M. M., Keegan, N., Short, M., Pillay, D., Manley, E., Cox, I. J., Heymann, D. L., Johnson, A. M., & McKendry, R. A. (2020). Digital technologies in the public-health response to COVID-19. *Nature Medicine, 26*(8), 1183–1192. Dostupné z: https://doi.org/10.1038/s41591-020-1011-4

CENELEC (2024) *úvodní stranka.* Citováno 9. ledna 2024. Dostupné z: https://www.cencenelec.eu/

Cave, S. (2020). The Problem with Intelligence: Its Value-Laden History and the Future of AI. In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* February 2020 (s. 29–35). New York: Association for Computing Machinery https://doi.org/10.1145/3375627.3375813

Charlesworth, Matthew. (2005). *An investigation of an undergraduate course module on the ethical aspects of information systems*. Citováno 1. března 2024. Dostupné z: https://core.ac.uk/download/pdf/145043415.pdf

City of Amsterdam Algorithm Register (n.d.). *What is the Algorithm Register?* Citováno 22. března 2024. Dostupné z: https://algoritmeregister.amsterdam.nl/en/ai-register/

City of Helsinki AI Register (n.d.). *What is AI Register?* Citováno 22. března 2024. Dostupné z: https://ai.hel.fi/en/ai-register/

Clayton, A. (2020, October 27). *How eugenics shaped statistics,* Nautilus. Citováno 20. února 2024. Dostupné z: https://nautil.us/how-eugenics-shaped-statistics-238014/

Cobbe, J. (2022 November 21) TECHNOCHAUVINISM Přednáška na univerzitě SciencesPo. Citováno 10. Ledna 2024. Dostupné z: https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2022/06/Jennifer-Cobbe-TECHNOCHAUVINISM-Policy-Brief-.pdf

Coeckelbergh, M. (2022). *The Political Philosophy of AI. An Introduction*. Cambridge: Polity Press.

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, Massachusetts: MIT Press.

Coeckelbergh, M. (2023). *Etika umělé inteligence*. Praha: Filosofia.

Coeckelbergh, M. & Reijers, W (2020). *Narrative and Technology Ethics*. Switzerland: Springer Nature.

Corbyn, Z. (2023, March 26). *AI expert Meredith Broussard: 'Racism, sexism and ableism are systemic problems.'* The Guardian. Citováno 21. ledna 2024. Dostupné z: https://www.theguardian.com/technology/2023/mar/26/artificial-intelligence-meredith-broussard-more-than-a-glitch-racism-sexism-ableism

Committee on Responsible Science (CRS), Committee on Science, Engineering, Medicine, and Public Policy (2017). *Fostering Integrity in Research. A Consensus Study Report*. Washington, DC: The National Academies Press. Citováno 30. února 2023. Dostupné z: https://nap.nationalacademies.org/login.php?record_id=21896

Czexpats in Science (2021). Kulveit, J. Multidisciplinarita ve vědě“ (2021)[86] Citováno 15. března 2024. Dostupné z: https://www.youtube.com/watch?v=8R67Sn8yaA0

Česká asociace umělé inteligence (2023, November 21). *Slovník pojmů - Česká asociace umělé inteligence*. Citováno 30. prosince 2023. Dostupné z: https://asociace.ai/slovnik-pojmu/

Česko. (1998). *Zákon č. 111/1998 Sb.Zákon o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách).* Citováno 3. Unora 2024. Dostupné z: https://www.e-sbirka.cz/sb/1998/111?zalozka=text

Česko. *(2002). Zákon č. 130/2002 Sb.Zákon o podpoře výzkumu a vývoje z veřejných prostředků a o změně některých souvisejících zákonů (zákon o podpoře výzkumu a vývoje).* Citováno 30. ledna 2024. Dostupné z: https://www.zakonyprolidi.cz/cs/2002-130

Česko (2020). *Sdělení č. 30/2020 Sb. m. s. Sdělení Ministerstva zahraničních věcí o sjednání Dodatkového protokolu k Úmluvě o lidských právech a biomedicíně souvisejícího s biomedicínským výzkumem.* Dostupné z: https://www.zakonyprolidi.cz/ms/2020-30

CoARA - Coalition for Advancing Research Assessment (Koalice pro reformu hodnocení výzkumu) (2022). *Agreement on reforming Research Assessment*. Citováno 29. ledna 2024. Dostupné z: https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf

Delloite (May 2023). *Perspective on New York City local law 144-21 and preparation for bias audits*. Citováno 12. února 2024. Dostupné z: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/audit/us-audit-nyc-hiring-law.pdf

Digitální a Informační Agentura (n.d.). *Budoucnost je digitální.* Citováno 6. února 2024. Dostupné z: https://www.dia.gov.cz/

Dignum, V. (2019). *Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way*. Switzerland: Springer Nature.

Drozenová, W. et al. (2010). *Etika vědy v České republice od historických kořenů k současné bioetice.* Praha: Filosofia.

Dzurilla, V., Očko, P. et al. (2019). *Digitální ekonomika a společnost. Vládní program digitalizace České republiky 2018+,* Úřad vlády ČR. Citováno dne 15. března

[86] https://www.youtube.com/watch?v=8R67Sn8yaA0

2023. Dostupné z: https://vlada.gov.cz/assets/ppov/rvis/zapisy_rvis/DES-2023.pdf

ENISA (2024). *About ENISA - The European Union Agency for Cybersecurity. Towards a Trusted and Cyber Secure Europe*. Citováno 19. února 2024. Dostupné z: https://www.enisa.europa.eu/about-enisa

ETHNA Systém (2023). *Deliverables.* Citováno 12. března 2024. Dostupné z: https://ethnasystem.eu/results/

EuroHPC JU (14. února 2024). *EuroHPC JU launches the procurement for LUMI-Q*. Citováno 20. února 2024. Dostupné z: https://eurohpc-ju.europa.eu/eurohpc-ju-launches-procurement-lumi-q-2024-02-14_en

European Commission (2021). *EU Grants. How to complete your ethics self-assessment version 2.0*. Citováno 20. března 2024. Dostupné z: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf

European Commission (2023). *The Digital Decade policy programme 2030. Factsheet*, Publications Office of the European Union. Citováno 28. února 2024. Dostupné z: https://digital-strategy.ec.europa.eu/en/library/policy-programme-path-digital-decade-factsheet

European Commission, Directorate-General for Justice and Consumers, EC DG JC (2019). *Liability for artificial intelligence and other emerging digital technologies,* Publications Office of the European Union. Citováno 28. února 2024. Dostupné z: https://data.europa.eu/doi/10.2838/573689

European Commission, Directorate-General for Research and Innovation, EC DG RI (2021). *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*. Citováno 28. února 2024. Dostupné z: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

European Commission, Directorate-General for Research and Innovation, EC DG RI (2022). *Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)*, Publications Office of the European Union. Citováno 28. února 2024. Dostupné z: https://data.europa.eu/doi/10.2777/935288

European Commission, Directorate-General for Research and Innovation, EC DG RI (n.d.). *European Group on Ethics in Science and New Technologies (EGE)*. Citováno 8. ledna 2024. Dostupné z: https://research-and-

innovation.ec.europa.eu/strategy/support-policy-making/scientific-support-eu-policies/european-group-ethics_en

European Commission, Directorate-General for Research and Innovation, EC DG RI (2024) *Living guidelines on the responsible use of generative AI in research*, Publications Office of the European Union. Citováno dne 25. března 2024. Dostupné z: https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf

European Commission, Horizon Europe Young Observer (2022). *List of selected candidates 5 December 2022.* Citováno 11. prosince 2023. Dostupné z: https://research-and-innovation.ec.europa.eu/document/download/ac06dbaa-11fe-4394-991f-08b7392058f1_en?filename=Young%20Observers%20list%202022%20v3.pdf

European Research Executive Agency (2023). *ERA Chairs.* Citováno 1. března 2024. Dostupné z: https://rea.ec.europa.eu/funding-and-grants/horizon-europe-widening-participation-and-spreading-excellence/era-chairs_en

European Society for Engineering Education (SEFI) (2024). *SEFI Conference 2024: Educating Responsible Engineers*. Citováno 19. března 2024. Dostupné z: https://sefi2024.eu/

Evropská komise (2024). Excelentní a důvěryhodná umělá intelligence. Citováno 28. března 2024. Dostupné z: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/excellence-and-trust-artificial-intelligence_cs

Evropská komise, High Level Expert Group on Artificial Intelligence (AI HLEG) (2019). *Etické pokyny pro důvěryhodnou umělou inteligenci*. Citováno 27. února 2024. Dostupné z. https://digital-strategy.ec.europa.eu/cs/library/ethics-guidelines-trustworthy-ai

Evans, K. D., Robbins, S. A., & Bryson, J. J. (2023). *Do We Collaborate With What We Design?* Topics in Cognitive Science. Citováno 30. ledna 2024. Dostupné z: https://doi.org/10.1111/tops.12682 s. 1-20

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI.* Berkman Klein Center for Internet & Society, 2020. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schäfer, B., Valcke, P., &

Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. Citováno 15. března 2024. Dostupné z: https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. (2023) *The Ethics of artificial Intelligence: Principles, Challenges, and Opportunities.* Oxford: Oxford University Press.

Future of Life Institute (FLI) (2024). *Article 56: Establishment and Structure of the European Artificial Intelligence Board.* Citováno 27. března 2024. Dostupné z: https://artificialintelligenceact.eu/article/56/

Future of Life Institute (FLI) (2017). Asilomar AI Principles. Citováno 6. února 2024. Dostupné z: https://futureoflife.org/open-letter/ai-principles/

Gasser, U., & Almeida, V.A. (2017). A Layered Model for AI Governance. *IEEE Internet Computing, 21*(6) (November): s.58–62. Citováno 10. ledna 2024. Dostupné z: doi:10.1109/mic.2017.4180835.

Glerup, C., & Horst, M. (2014). Mapping 'social responsibility' in science, *Journal of Responsible Innovation, 1*(1), 31–50. Citováno 10. ledna 2024. Dostupné z: https://doi.org/10.1080/23299460.2014.882077

Global Partnership on Artificial Intelligence, the (n.d.). *About GPAI*. Citováno 3. března 2024. Dostupné z: https://www.gpai.ai/about/

Gmyrek, P., Berg, J. & Bescond, D. (2023). *Generative AI and Jobs: A global analysis of potential effects on job quantity and quality. ILO Working Paper 96.* Geneva: International Labour Office.

González-Esteban, E., Feenstra, R. A., & Camarinha-Matos, L. M. (2023). *Ethics and responsible research and innovation in practice: The ETHNA System Project*. Springer Nature.

Green, A. (2023), "Artificial intelligence and jobs: No signs of slowing labour demand (yet)", in *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*, OECD Publishing, Paris, https://doi.org/10.1787/9c86de40-en.

Grygar, F., Zamarovský, P. (6.4.2023) *Filosofie a věda (EKK, AV ČR 6.4.2023) (v2).* Dostupné z: https://youtu.be/jbYkrdbmJ4c?si=Ui1GJGR7-kRVmbfZ&t=2699, osobní účast 6. 4. 2023.

Guha, R.V., Manyika, J. (2023). *Data Commons is using AI to make the world's public data more accessible and helpful*. Google. Citováno 15. srpna 2023. Dostupné z: https://blog.google/technology/ai/google-data-commons-ai/

Hanemaayer, A. (2022): Introduction: Critical Insights-Bringing the social sciences and humanities to AI. In Hanemaayer (Ed.). *Artificial Intelligence and Its Discontents, Social and Cultural Studies of Robots and* AI (1-20). Switzerland: Springer Nature.

Heaven, W. D. (2022, August 2). *Hundreds of AI tools have been built to catch covid. None of them helped.* MIT Technology Review. Citováno 11. března 2024. Dostupné z: https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

Ho et al. (2023). *International Institutions for Advanced AI*. Arxiv. Citováno 10. března 2023. Dostupné z: https://arxiv.org/pdf/2307.04699.pdf

Holzinger, A., Kieseberg, P., Weippl, E., & Tjoa, A. M. (2018). Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In *Springer eBooks* (s. 1–8). Springer Nature. https://doi.org/10.1007/978-3-319-99740-7_1 (s. 1-8)

Hosanagar, K. (2019). *A Human's Guide to Machine Intelligence: How Algorithms Are Shaping Our Lives and How We Can Stay in Control*. New York: Viking.

HC 1769, House of Commons UK (Poslanecká sněmovna), Science, Innovation and Technology Committee (2023, 19. července). *The governance of artificial intelligence: interim report, Ninth Report of Session 2022–23. Report HC 1769, together with formal minutes relating to the report*. Citováno 8. března 2024. Dostupné z: https://committees.parliament.uk/publications/41130/documents/205611/default/

HC 945, House of Commons UK (Poslanecká sněmovna), Science, Innovation and Technology Committee (2023, 22. února). Transcript z jednání. Citováno 8. března 2024. Dostupné z: https://committees.parliament.uk/oralevidence/12709/pdf/

Hrivňák, Tomáš (2023, July 26). Co dělat, když umělá inteligence už i programuje? Výzkumnice z DeepMind říká, jaké schopnosti budou potřeba i nadále. *Deník N*. Citováno 13 ledna 2024. Dostupné z: https://denikn.cz/1194578/co-delat-kdyz-umela-inteligence-uz-i-programuje-vyzkumnice-z-deepmind-rika-jake-schopnosti-budou-potreba-i-nadale/?ref=list

Hubálková, P. (2022, March 30). *Jan Kulveit: Zabýváme se výzkumem rizik, jež lidstvo podceňuje*. Vědavýzkum.cz. Citováno 30. ledna 2024. Dostupné z:

https://vedavyzkum.cz/rozhovory/rozhovory/jan-kulveit-zabyvame-se-vyzkumem-rizik-jez-lidstvo-podcenuje

Humancompatible.org (n.d.). *Let's make AI human-compatible!* Citováno 2. ledna 2024. Dostupné z: https://humancompatible.org/

Hutson, M. (2023). Rules to keep AI in check: nations carve different paths for tech regulation. *Nature, 620*(7973), s.260–263. Citováno 30. března 2024. Dostupné z: https://doi.org/10.1038/d41586-023-02491-y

Christian, B. (2020). *The alignment problem: Machine Learning and Human Values*. National Geographic Books.

International Conference of Social Robotics (2022). *Eco-socio-botics 2022: Social Robotics for Sustainability*. Citováno 31. března 2024. Dostupné z: https://sites.google.com/view/ecosociobotics2022/home

Kaufmann, J.C., (2010). *Chápající Rozhovor*. Praha: SLON.

Kempelen Institute of Intelligent Technologies (n.d.). *Harnessing research for humans & industries. Discover Kinit.* Citováno 21. března 2024. Dostupné z: https://kinit.sk/

Kolaříková, L. & Horák, F. (2020). *Umělá inteligence & právo.* Praha: Wolters Kluwer.

Korinek, A. (2020). Integrating Ethical Values and Economic Value to Steer Progress in Artificial Intelligence. In Dubber, M. D., Pasquale, F., & Das, S. (Eds.). *The Oxford Handbook of Ethics of AI* (s. 475-491). New York: Oxford University Press.

Krausová, A. et al (2018) *Výzkum potenciálnu rozvoje umělé inteligence v České republice. Analýza právně-etických aspektů rozvoje umělé inteligence a jejích aplikací v ČR.* TAČR, TI00UVCR001.

Kretser, A., Murphy, D., Bertuzzi, S. et al. (2019) Scientific Integrity Principles and Best Practices: Recommendations from a Scientific Integrity Consortium. *Science and Engineering Ethics 25*, s.327–355. https://doi.org/10.1007/s11948-019-00094-3

Lee, C. (2017). *James Damore has sued Google. His infamous memo on women in tech is still nonsense.* Vox. Citováno 31. října 2019. Dostupné z: https://www.vox.com/the-big-idea/2017/8/11/16130452/google-memo-women-tech-biology-sexism.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector v1*. Zenodo. https://doi.org/10.5281/zenodo.3240529

Luke, J., Porter, D. & Santhanam, P. (2022). *Beyond Algorithms, Delivering AI for Business*. London, CRC Press.

Maguire, Y. (Aug 28, 2023). *New sustainability tools help businesses and cities map environmental information*. Google. Citováno 3. ledna 2024. Dostupné z: https://blog.google/products/maps/google-maps-apis-environment-sustainability/

Marušáková (2023). *Renata Veselská: Etika výzkumu je z podstaty mezioborová záležitost*. Universitas. Citováno 10. března 2023. Dostupné z: https://www.universitas.cz/tema/10006-veselska-etika-vyzkumu-je-z-podstaty-mezioborova-zalezitost

Mařík, V., Štěpánková, O., & Lažanský, J. (1993). *Umělá inteligence* (1). Praha: Academia.

Masaryk University (2023). *GRANTS WEEK: Your opportunity to find out all you need to know about international research funding. Programme*. Citováno 22. března 2024. Dostupné z: https://www.grantsweek.muni.cz/programme

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, L., Lyons, T., Manyika, J., Ngo, H., Niebles, J.C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). *The AI Index 2023 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. Citováno 13. března 2024. Dostupné z: https://aiindex.stanford.edu/report/

MOOC center, University of Helsinki (n.d.). *The Ethics of AI*. Citováno 12. ledna 2024. Dostupné z: https://ethics-of-ai.mooc.fi/

MPO (2024). *Informace o plnění a aktualizaci Národní strategie umělé inteligence v České republice*. Citováno 22. ledna 2024. Dostupné z. https://www.mpo.cz/assets/cz/rozcestnik/ministerstvo/aplikace-zakona-c-106-1999-sb/informace-zverejnovane-podle-paragrafu-5-odstavec-3-zakona/2024/3/Informace-o-plneni-a-aktualizaci-NAIS.pdf

National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Committee on Responsible Science (2017). *Fostering Integrity in Research*. Washington (DC): National Academies Press (US).

NATO (2021). *An Artificial Intelligence Strategy for NATO*. Citováno 10. března 2024. Dostupné z: https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.

NATO (2020). *Science & Technology Trends 2020-2040.* Citováno 10. března 2024. Dostupné z: https://www.nato.int/nato_static_fl2014/assets/pdf/2020/4/pdf/190422-ST_Tech_Trends_Report_2020-2040.pdf

Nature Editorial (2023). Stop talking about tomorrow's AI doomsday when AI poses risks today. *Nature*, *618*(7967), 885–886. https://doi.org/10.1038/d41586-023-02094-7

NeurIPS (2024). *NeurIPS Code of Ethics*. Citováno 28. března 2024. Dostupné z: https://neurips.cc/public/EthicsGuidelines

NIST (January 2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Citováno 7. listopadu 2023. Dostupné z: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

NÚKIB Národní úřad pro kybernetickou a informační bezpečnost. (n.d.) *Úvodní stránka.* Citováno 11. března 2024. Dostupné z: https://nukib.gov.cz/

OECD (n.d.). *Technology Governance*. Citováno 12. února 2024. Dostupné z: https://www.oecd.org/sti/science-technology-innovation-outlook/technology-governance/

OECD (2023a). *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*. OECD Publishing, Paris. Dostupné z. https://doi.org/10.1787/08785bba-en

OECD (2023b). *OECD Science, Technology and Innovation Outlook 2023: Enabling Transitions in Times of Disruption*. OECD Publishing, Paris. Dostupné z: https://doi.org/10.1787/0b55736e-en.

OECD.AI (2023). *Assessment for Responsible Artificial Intelligence.* OECD.AI. Citováno 12 února 2024. Dostupné z: https://oecd.ai/en/catalogue/tools/z-inspection/tool-use-cases/assessment-for-responsible-artificial-intelligence.

OECD.AI, (2024). *What is AI? Can you make a clear distinction between AI and non-AI systems?* OECD.AI. Citováno 11. března 2024. Dostupné z: https://oecd.ai/en/wonk/definition,

OSN, UN, (n.d.). *Cíle udržitelného rozvoje (SDGs)*. Citováno 25. února 2024. Dostupné z: https://osn.cz/osn/hlavni-temata/cile-udrzitelneho-rozvoje-sdgs/

Perrigo, B. (2022, January 18). Why Timnit Gebru Isn't Waiting for Big Tech to Fix AI's Problems. *Time*. Citováno 9. Března 2024. Dostupné z https://time.com/6132399/timnit-gebru-ai-google/

Province of Fryslân, Rijks ICT Gilde & the Z-Inspection® Initiative (2023). *Assessing the trustworthiness of an AI system in practice. Lessons Learned from the expert examination of the AI system "Monitoring grassification of heather fields"*. Citováno 12. února 2024. Dostupné z: https://www.rijksorganisatieodi.nl/binaries/rijksorganisatieodi/documenten/publicaties/2023/08/01/the-main-lessons-learned-pilot-ai-systeem/Summary+Z-Inspection+pilot+Netherlands+-+Main+lessons+learned.pdf

PRG.AI, (n.d.). *Artificial intelligence and human rights: risks, opportunities and regulation*. Interdisciplinary research project funded by the Technology Agency of the Czech Republic (2021-2023) No. TL05000484. Citováno 30. března 2024. Dostupné z: https://prg.ai/projekty/ai-lidska-prava/

RAND (2024). *Is AI an Existential Risk? Q&A with RAND Experts*. Citováno 10. ledna 2024. Dostupné z: https://www.rand.org/pubs/commentary/2024/03/is-ai-an-existential-risk-qa-with-rand-experts.html

Rauhala, M., (2023). *Caring rather than Clearing: Introducing the TU Wien Pilot Research Ethics Committee*. EARMA Conference Prague 2023. Citováno 26. dubna 2023. Dostupné z: https://earma.org/abstracts/submission/642/view/

Resnik, D. B., & Elliott, K. C. (2015). The ethical challenges of socially responsible science. Accountability in *Research, 23*(1), 31–46. Dostupné z: https://doi.org/10.1080/08989621.2014.1002608

Richta, R. (1969). *Civilizace na rozcestí: společenské a lidské souvislosti vědeckotechnické revoluce*. Praha: Svoboda.

Roberts, M. et al, (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence, 3*(3), s.199–217. https://doi.org/10.1038/s42256-021-00307-0

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach, 3rd ed*. Pearson Education

RRI Tools (2016). *Welcome to the RRI Toolkit: Towards an open science and innovation system that tackles the societal challenges of our world*. Citováno 21. března 2024. Dostupné z: https://rri-tools.eu/

RVVI, ÚV ČR (2019). Inovační strategie České republiky 2019–2030. Citováno 14. března 2024. Dostupné z: https://vyzkum.gov.cz/FrontClanek.aspx?idsekce=866015

Řehořek, T. & Surynek, P. (2023). Přednáškové materiály k předmětu Základy umělé inteligence BI-ZUM na FIT ČVUT, který vede doc. RNDr. Pavel Surynek. Autorem je Ing. Tomáš Řehořek.

Saltz J. (October 6, 2023). *What is the AI Life Cycle?* Datascience. Citováno 30. Ledna 2024. Dostupné z: https://www.datascience-pm.com/ai-lifecycle/

Schwartz, R. et al. (2022, March) *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* National Institute of Standards and Technology (NIST), USA. Citováno 23 února 2024. Dostupné z: https://acresia.com/images/Documents/Novinky/dokumenty/Towards-Standard-for-Identifying-and-Managing-Bias-in-Artificial-Intelligence.pdf

SHERPA consortium (n.d.) *Shaping the Ethical Dimensions of Smart Information Systems: A European Perspective*. Citováno 14. března 2024. Dostupné z: https://www.project-sherpa.eu/

Sienna project (n.d.) *SIENNA: Technology, ethics and human rights.* Citováno 14. března 2024. Dostupné z: https://www.sienna-project.eu/

Snow, C. P. (2012). *The two cultures*. Cambridge University Press.

Srikumar, M., Finlay, R., Abuhamad, G. et al. (2022). Advancing ethics review practices in AI research. *Nature Machine Intelligence 4*, s.1061–1064 (2022). https://doi.org/10.1038/s42256-022-00585-2

Státní úřad inspekce práce (2024). *Akt o umělé inteligenci (AI Act) schválen. Klíčový průvodce novou regulací EU*. Citováno 21. března 2024. Dostupné z: https://www.bezpecnostprace.info/umela-inteligence-ai/ai-act/

Steen, M. (2022). *Ethics for people who work in tech.* Oxon: Chapman & Hall/CRC.

Šabanović, S. et al. (2023) "Robots for good": Ten defining questions. *Science Robotics* 8, eadl4238 DOI:10.1126/scirobotics.adl4238

Šmuclerová M., Král, L., Drchal, J. et al. (2023a). *Umělá inteligence a lidská práva: rizika, příležitosti a regulace*. Souhrnná výzkumná zpráva, TAČR č. TL05000484.

Šmuclerová M., Král, L., Drchal, J. et al. (2023b). *Umělá inteligence a lidská práva: rizika, příležitosti a regulace životního cyklu AI, Policy paper*. Policy paper pro veřejnou správu. TAČR č. TL05000484.

Šmuclerová M., Král, L., Drchal, J. et al. (2023c). *Umělá inteligence a lidská práva: Soubor doporučení pro subjekty životního cyklu AI*, Policy paper.

Šmuclerová M., Král, L., Drchal, J. et al. (2023, 21. 9.) *Umělá inteligence a lidská práva: rizika, příležitosti a regulace*. Závěrečný workshop projektu, 21. 9. 2023, ČVUT.

Tuples.ai (2023) *Building trustworthy planning and scheduling systems*. Citováno 2. ledna 2024. Dostupné z: https://tuples.ai/

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. Citováno 11. srpna 2023. Dostupné z: http://www.jstor.org/stable/2251299

UNESCO (2023). *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*. SHS/REI/BIO/REC-AIETHICS-TOOL-EIA/2023. Dostupné z: https://unesdoc.unesco.org/ark:/48223/pf0000386276.locale=en

UNESCO (2023). *Recommendation on the Ethics of Artificial Intelligence*. SHS/2023/PI/H/5. Dostupné z: https://unesdoc.unesco.org/ark:/48223/pf0000386510.locale=en

Úřad vlády ČR (n.d.). *Základní pojmy výzkumu a vývoje v OECD a EU*. Citováno 22. ledna 2024. Dostupné z: https://vyzkum.gov.cz/FrontClanek.aspx?idsekce=932

Úřad vlády ČR (2019). *Národní strategie umělé inteligence ČR (NAIS)*. Citováno dne 15. března 2023. Dostupné z: https://www.mpo.cz/assets/cz/podnikani/2023/1/NAIS_kveten_2019.pdf

Vantard, M., Galland, C., Knoop, M. (2023). Interdisciplinary research: Motivations and challenges for researcher careers. *Quantitative Science Studies* 2023; 4 (3): 711–727. doi: https://doi.org/10.1162/qss_a_00265

Veselská, R., Širůček, J., Kuře, J. & Šerek, J. (2023). *Trasa 8 – Etika výzkumu (Research Ethics)*. Citováno 15. března 2024. Dostupné z: https://is.muni.cz/do/rect/el/estud/lf/js23/vedecka_integrita/web/pages/trasa_8.htmlů

Vláda ČR (n.d.). *Odbor věcných politik EU*. Citováno 5. února 2024. Dostupné z: https://vlada.gov.cz/cz/evropske-zalezitosti/organizace-utvaru/koordinace-ru%20stovych-politik/uvod-119953/

West, S. L. (2017). Data Capitalism: Redefining the Logics of Surveillance and Privacy. *Business & Society*, *58*(1), 20–41. https://doi.org/10.1177/0007650317718185.

White House, the (July 2022). *U.S. and U.K. Launch Innovation Prize Challenges in Privacy-Enhancing Technologies to Tackle Financial Crime and Public Health Emergencies.* Citováno 17. února 2024. Dostupné z: https://www.whitehouse.gov/ostp/news-updates/2022/07/20/u-s-and-u-k-launch-innovation-prize-challenges-in-privacy-enhancing-technologies-to-tackle-financial-crime-and-public-health-emergencies/

White House, the (July 2023). *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*. Citováno 5. února 2024. Dostupné z: https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* London: Nuffield Foundation.

Wiener, N. (1960). Some Moral and Technical Consequences of Automation. *Science*, *131*(3410), s.1355–1358. http://www.jstor.org/stable/1705998

Zuboff, S. (2022). *Věk kapitalismu dohledu: boj o budoucnost lidstva u nové hranice moci.* Praha: Argo

Seznam zkratek

AI	umělá inteligence z angl. artificial intelligence
HE	Horizon Europe
NAIS	Národní strategie umělé inteligence ČR
OECD	Organizace pro hospodářskou spolupráci a rozvoj (Organisation for Economic Co-operation and Development)
OSN	Organizace spojených národů
RAI	Odpovědná umělá inteligence (Responsible AI)
RIS3	Národní výzkumná a inovační strategie pro inteligentní specializaci (Research and Innovation Strategy for Smart Specialisation)
RMAs	Research Managers and Administrators (manažeři a administrátoři výzkumných projektů)
RRI	Responsible Research and Innovation
SSH	zkratka pro společenskovědní a humanitní obory (z angl. social sciences and humanities), někdy se používají jiné zkratky, např. SHAPE (angl. Social Sciences, Humanities and the Arts for People and the Economy, česky potom SHU
STEM	zkratka pro obory z oblasti přírodních věd (Science), technologií (Technology), techniky (Engineering) a matematiky (Mathematics)
UI	Umělá inteligence (někde také jako AI – např. AI Akt).
VaVaI	Výzkum, vývoj a inovace
XAI	Explainable AI (vysvětlitelná UI) – prevence fenoménu black box

PŘÍLOHA

Seznam obrázků