Záverečná výskumná správa o modelovaní morálnej citlivosti tvorcov systémov AI

Názov projektu	**Morálna citlivosť a ľudské práva pre spracovanie jazykov s obmedzenými zdrojmi**
Akronym projektu	**SensAI**
Kód projektu	**09I01-03-V04-00100**
Začiatok projektu	**01.01.2025**
Trvanie projektu	**18 mesiacov**

Financované Európskou úniou NextGenerationEU

PLÁN OBNOVY

Sumarizácia

Výskumná správa sumarizuje výsledky riešenia úlohy U3.3 zameranej na výskum modelov morálnej citlivosti tvorcov systémov umelej inteligencie. Cieľom výskumu bolo preskúmať existujúce konceptualizácie morálnej citlivosti, analyzovať dostupné prístupy k jej meraniu a identifikovať metodologické východiská pre ich adaptáciu do prostredia vývoja a nasadzovania systémov AI.

V rámci výskumu bol vytvorený teoretický rámec širokej morálnej citlivosti, identifikované špecifiká AI praxe relevantné pre jej meranie a systematicky analyzované existujúce metodiky využívané v rôznych profesijných doménach. Na základe tejto analýzy boli identifikované metodologické princípy a najvhodnejšie prístupy pre budúcu adaptáciu do AI prostredia, pričom ako najperspektívnejší základ bol vyhodnotený Engineering Ethical Reasoning Instrument (EERI). Hlavným výsledkom výskumu je vytvorenie konceptuálneho a metodologického rámca, ktorý predstavuje východisko pre ďalší vývoj empirických nástrojov na meranie morálnej citlivosti tvorcov systémov AI.

1. Kontext, východiská a vzťah k projektu ALFIE

Táto výskumná časť projektu SensAI priamo nadväzuje na výskumné a vývojové aktivity realizované v rámci projektu ALFIE (Assessment of Learning Technologies and Frameworks for Intelligent and Ethical AI), a to hlavne na úlohu *T2.2 – Analysis of ethical principles in modern AI solutions*, ktorú KInIT v rámci projektu ALFIE vedie.

Cieľom tejto úlohy je komplexne analyzovať existujúce etické princípy, rámce a smernice pre dôveryhodnú AI, identifikovať ich silné a slabé stránky a hodnotiť ich praktickú uplatniteľnosť v reálnych kontextoch vývoja a nasadzovania AI systémov. V rámci tejto analýzy je osobitný dôraz kladený na skúmanie účinnosti etických princípov pri znižovaní biasov, zvyšovaní férovosti, transparentnosti, zodpovednosti a morálnej citlivosti poskytovateľov AI. Pôvodná výskumná úloha pracuje s predpokladom, že na to aby boli takéto nástroje efektívne a účinné, mali by podporovať schopnosti tvorcov AI systémov rozpoznávať, interpertovať a zohľadnovať morálne aspekty pri výskume, vývoji a nasadtovaní systémov umelej inteligencie.

Projekt SensAI sa preto zameriava na hlbšie konceptuálne, empirické a metodologické uchopenie pojmu morálnej citlivosti, a to špecificky v kontexte vývoja dôveryhodnej umelej inteligencie. Cieľom je adresovať medzeru medzi abstraktnými etickými princípami a ich operacionalizaciou do konkrétnych praktík. Okrem toho je ambíciou projektu SensAI posunúť poznatky z výskumných úloh ALFIE smerom k praktickému overeniu, ako môžu etické rámce a odporúčania pre dôveryhodnú umelú inteligenciu reálne prispievať k zodpovednejšiemu vývoju umelej inteligencie.

2. Priebeh prác a aktivít

Výskum v rámci úlohy U3.3 prebiehal v dvoch vzájomne nadväzujúcich fázach. V prvej fáze sa výskumný tím zameral na konceptuálne a metodologické zmapovanie problematiky morálnej citlivosti. Cieľom bolo identifikovať existujúce prístupy ku konceptualizácii morálnej citlivosti, preskúmať jej vzťah k príbuzným pojmom, najmä morálnej kompetencii, a analyzovať dostupné metódy jej merania naprieč rôznymi profesijnými oblasťami.

Na základe systematického prieskumu literatúry bola formulovaná pracovná koncepcia morálnej citlivosti a vytvorený analytický rámec siedmich komponentov morálnej citlivosti relevantných pre kontext vývoja a nasadzovania systémov umelej inteligencie. Súčasne boli analyzované existujúce nástroje na meranie morálnej citlivosti a príbuzných konštruktov v oblastiach ako zdravotníctvo, vzdelávanie, manažment a podnikateľská etika.

Výsledky prvej fázy ukázali, že napriek existencii viacerých konceptualizácií a meracích nástrojov zostáva otvorenou otázkou teoretické postavenie morálnej citlivosti vo vzťahu k morálnej kompetencii, praktickej múdrosti (phronesis) a širším wisdom tradíciám. Zároveň sa ukázalo, že existujúce nástroje boli vytvorené pre odlišné profesijné prostredia a nemožno ich priamo preniesť do kontextu vývoja umelej inteligencie.

Tieto zistenia vytvorili východisko pre druhú fázu výskumu, ktorá sa zamerala na hlbšie teoretické ukotvenie konceptu morálnej citlivosti, analýzu jeho vzťahu k modelom praktickej

múdrosti a morálnej kompetencie, formuláciu predpokladov pre meranie morálnej citlivosti v technologickom prostredí a identifikáciu vhodných metodologických prístupov pre budúce empirické skúmanie morálnej citlivosti tvorcov systémov AI.

3. Teoretické ukotvenie konceptu morálnej citlivosti

Jedným z hlavných zistení prvej fázy výskumu bolo, že napriek rastúcemu záujmu o morálnu citlivosť v kontexte umelej inteligencie zostáva nejasné jej teoretické postavenie vo vzťahu k príbuzným konceptom, ako sú morálna kompetencia, praktická múdrosť (phronesis), etická expertíza alebo širšie tradície výskumu múdrosti. Táto nejasnosť predstavuje významný problém aj z metodologického hľadiska. Bez jasného vymedzenia toho, aký typ schopnosti morálna citlivosť predstavuje, je totiž náročné určiť, ktoré aspekty by mali byť predmetom merania a akým spôsobom by mali byť operacionalizované.

Z tohto dôvodu sa druhá fáza výskumu zamerala na hlbšie teoretické ukotvenie konceptu Broad Moral Sensitivity (BMS), ktorý bol vypracovaný v predchádzajúcich etapách projektu. Výsledkom tejto práce bol vedecký článok “Ethics of AI Practitioners: Positioning a Broader Conception of Moral Sensitivity within the Virtue and Wisdom Traditions”, ktorého cieľom bolo preskúmať vzťah medzi BMS a vybranými modelmi praktickej múdrosti a etickej expertízy a objasniť, či možno morálnu citlivosť chápať ako samostatnú kompetenciu relevantnú pre AI prax.

V článku "Ethics of AI Practitioners: Positioning a Broader Conception of Moral Sensitivity within the Virtue and Wisdom Traditions" sme sa zaoberali otázkou, ako možno širšiu koncepciu morálnej citlivosti – broad moral sensitivity (BMS) – zaradiť do súčasných diskusií o etike umelej inteligencie, praktickej múdrosti (*phronesis*) a etike cností. Východiskom textu je presvedčenie, že AI etika sa nemôže uspokojiť iba s abstraktnými etickými princípmi. Hoci princípy ako spravodlivosť, transparentnosť, zodpovednosť alebo rešpektovanie autonómie zohrávajú dôležitú úlohu, samy osebe ešte nezaručujú, že budú v praxi správne pochopené, uplatnené a premietnuté do konkrétnych rozhodnutí AI vývojárov.

Vychádzame zo širšieho posunu v AI etike, ktorým je posun od formulovania abstraktných princípov k skúmaniu praktických podmienok ich implementácie. Kritizovaný je najmä dominantný deontologický alebo principlistický prístup, ktorý síce ponúka dôležité normatívne usmernenia, no často zostáva príliš všeobecný, málo citlivý na kontext a nedostatočne účinný pri reálnom rozhodovaní. Problémom nie je iba to, že princípy sú abstraktné, ale aj to, že ich uplatnenie vyžaduje ďalšie schopnosti: rozpoznať, že situácia je morálne relevantná, správne ju interpretovať, pochopiť jej dosah na rôznych aktérov, predvídať možné dôsledky a vedieť odôvodniť, čo by sa malo urobiť.

Práve tu vstupuje do hry BMS. Úzke chápanie redukuje morálnu citlivosť najmä na morálne vnímanie/uvedomenie – teda schopnosť zaregistrovať, že určitá situácia má morálny rozmer. Širšie chápanie je však bohatšie. BMS zahŕňa viacero komponentov: morálne

vnímanie/uvedomenie, prisúdenie dôležitosti identifikovanému problému, morálne rámcovanie situácie, zaujatie perspektívy (z pozície iných aktérov), riešenie hodnotových konfliktov, predvídavosť a napokon úsudok o tom, čo by sa malo urobiť. Takto širšie chápaná morálna citlivosť nie je iba prvým krokom k etickému konaniu, ale komplexnou schopnosťou orientovať sa v morálne neprehľadných, technicky zložitých a sociálne rozptýlených situáciách AI praxe.

Cieľom článku však nie je predstaviť BMS ako nový zoznam schopností, ale pýtame sa, ako tento koncept súvisí s tromi etablovanými modelmi praktickej múdrosti, ktoré sú vnímané ako odpoveď na spomínanú priepasť medzi morálnymi princípmi a morálnym konaním. Konkrétne ide o Neo-Aristotelian *Phronesis* Model (APM), Aretai Centre Model of *Phronesis* (ACMP) a Common Wisdom Model (CWM) a jedna z našich hlavných otázok znie, či je BMS len novým názvom pre praktickú múdrosť, alebo ide o samostatnú morálnu kompetenciu?

APM chápe praktickú múdrosť ako viac-komponentnú schopnosť, ktorá prepája morálne videnie konkrétnej situácie s celkovou predstavou dobrého života a s konaním. Má štyri základné funkcie. Prvou je konštruktívna funkcia, teda schopnosť rozpoznať morálne relevantné prvky situácie. Druhou je regulácia emócií, čiže schopnosť rozumovo formovať emocionálne reakcie. Treťou je tzv. morálna identita (blueprint), teda celkový morálny obraz alebo vnútorný kompas človeka. Štvrtou je integračná funkcia (morálne rozhodovanie), ktorá pomáha rozhodovať medzi rôznymi cnosťami alebo hodnotami v konkrétnej situácii.

Medzi APM a BMS existujú zjavné prieniky. Najsilnejší je medzi konštruktívnou funkciou *phronesis* a komponentom morálneho uvedomenia v BMS, keďže obe zdôrazňujú, že človek musí najprv vidieť situáciu ako morálne významnú. Slabšie podobnosti možno nájsť aj medzi reguláciou emócií a zaujímaním perspektívy iných aktérov, prípadne medzi morálnou identitou a rámcovaním v BMS. Napriek tomu sa BMS nedá jednoducho stotožniť s neo-Aristotelovskou *phronesis*, keďže *phronesis* smeruje celkovému formovaniu ľudského charakteru. BMS je v tomto kontexte užšia a viac zameraná na štruktúrovanie morálneho uvažovania v profesionálnej AI praxi. Nepokrýva celistvú prestavu dobrého života ani dlhodobú kultiváciu charakteru.

Na rozdiel od APM, ACMP sa odkláňa od aristotelovského rámca a *phronesis* nechápe primárne ako cnosť charakteru, ale ako etickú expertízu alebo zručnosť. Autori tohto modelu hovoria o "virtue molecularism" – teda o predstave, že jednotlivé cnosti nie sú samostatné atómy koordinované *phronesis*, ale prejavy jednej zjednotenej etickej expertízy. Cieľom nie je budovanie charakteru človeka, ale cieľom je eticky kompetentný expert, ktorý má morálne poznanie, citlivosť na konkrétny kontext, afektívnu orientáciu na dobré ciele a zároveň otvorenosť voči novým situáciám a vlastným limitom.
Aj tu možno nájsť prieniky s BMS. Morálne uvedomenie v BMS sa podobá jemne vyladenej morálnej percepcii v ACMP a zaujímanie perspektívy iných aktérov možno čiastočne spojiť s afektívnou orientáciou etického experta. Ďalšie komponenty BMS – napríklad rámcovanie, riešenie hodnotových či úsudok – možno chápať ako prejavy metakognitívneho morálneho poznania, s ktorým tento model pracuje. Rozdiel je však zásadný. ACMP zdôrazňuje jednotnú, celostnú etickú expertízu, zatiaľ čo BMS pracuje s rozlíšenými komponentmi, ktoré majú AI praktici vedieť identifikovať a rozvíjať.

Do tretice, CWM nevychádza z cnostnej etiky, ale zo psychologického skúmania toho, ako ľudia morálne ukotveným spôsobom spracúvajú sociálne a kognitívne náročné situácie. Koncepcia múdrosti má dva hlavné piliere: perspektívnu metakogníciu a morálne ašpirácie. Perspektívna metakognícia zahŕňa schopnosť uvažovať o situácii z viacerých uhlov, uvedomovať si limity vlastného poznania, zvažovať rôzne záujmy a rozumieť komplexnosti konkrétnej situácie. Morálne ašpirácie zahŕňajú orientáciu na pravdu, vyvažovanie vlastných a cudzích záujmov a zameranie na spoločnú ľudskosť.

Z nášho skúmania sa BMS javí byť najbližšie práve tomuto modelu, keďže mnohé komponenty BMS možno chápať ako formy metakognitívnych schopností, a obe sa sústreďujú skôr na reflexívne a interpretačné schopnosti než na plné formovanie charakteru. Stále tu ale nemôžeme hovoriť o totožnosti medzi CWM a BMS: CWM je všeobecný model múdrosti, zatiaľ čo BMS je doménovo špecifický koncept vypracovaný pre AI prax, a BMS navyše výraznejšie zohľadňuje špecifiká sociotechnických systémov, ako sú distribuovaná zodpovednosť, skupinová dynamika, technická komplexnosť či dlhodobé a nepriame dôsledky AI systémov.

V závere článku tak zdôrazňujeme, že BMS by sa nemala chápať ani ako plná cnosť charakteru v aristotelovskom zmysle, ani ako kompletná etická expertíza. Skôr ide o samostatnú, praxovo orientovanú kompetenciu etického uvažovania, ktorá pripravuje pôdu pre zodpovedný konanie v AI praxi. BMS pomáha AI vývojárom lepšie vidieť, interpretovať a hodnotiť morálne relevantné aspekty ich práce, no sama ešte nezaručuje správne konanie. S tým súvisí aj ďalší prínos článku, ktorým je to, že BMS môžeme koncepčne chápať ako most medzi abstraktnými princípmi AI etiky a konkrétnou profesionálnou praxou. Ak sa AI etika nemá zredukovať na akýsi checkbox, teda formálne odškrtávanie princípov, potrebuje rozvíjať schopnosti ľudí, ktorí AI systémy navrhujú a nasadzujú. BMS pritom ponúka praktický rámec pre etické vzdelávanie AI profesionálov – napríklad cez prípadové štúdie, mapovanie stakeholderov, scenárové uvažovanie, identifikáciu rizík a reflexiu hodnotových konfliktov. Článok tak prispieva k širšej debate o tom, ako posunúť AI etiku z roviny princípov do roviny morálne zodpovednej praxe.

4. Vybrané empirické prístupy k meraniu morálnych schopností

Hoci koncept Broad Moral Sensitivity poskytuje teoretický rámec pre pochopenie morálneho uvažovania AI profesionálov, návrh budúceho meracieho nástroja si vyžaduje aj analýzu existujúcich empirických metodík. V ďalšej fáze výskumu sme preto preskúmali vybrané nástroje určené na meranie morálnych schopností, pričom osobitnú pozornosť sme venovali dvom nástrojom, konkrétne Short Phronesis Measure (SPM) a Moral Competence Test (MCT) Georga Linda (2019). Cieľom nebolo posúdiť jeho priamu aplikovateľnosť v AI praxi, ale identifikovať metodologické princípy, ktoré by mohli byť relevantné pri budúcej adaptácii nástrojov na meranie morálnej citlivosti v technologických profesiách.

Analýza bola vedená s cieľom identifikovať metodologické princípy, ktoré by mohli byť využité pri návrhu budúceho nástroja pre AI prax. Nezameriavali sme sa preto iba na psychometrické vlastnosti jednotlivých testov, ale predovšetkým na to, aké typy morálnych situácií zachytávajú, aké schopnosti merajú a aké metodologické prvky využívajú na operacionalizáciu morálneho uvažovania. V nasledujúcej časti preto stručne predstavujeme oba vybrané prístupy a diskutujeme ich potenciál z pohľadu budúcej adaptácie do prostredia vývoja systémov umelej inteligencie.

Autori tohto modelu (McLoughlin et al., 2025) vyvinuli a validovali aj empirický nástroj na meranie takto chápanej *phronesis* a nazvali ho Short *Phronesis* Measure (SPM) . Ide o skrátený dotazníkový nástroj (vyplnenie trvá približne 15 – 20 minút), ktorý má byť praktickejšou alternatívou k predchádzajúcemu približne 45-minútovému Long *Phronesis* Measure (približná dĺžka 45 minút). SPM kombinuje objektívne úlohy so sebahodnotiacimi položkami. Pri morálnom vnímaní respondenti posudzujú krátke scenáre: najprv určujú, či daná situácia vôbec má morálny dosah na charakter človeka, a následne vyberajú, ktoré cnosti sú v danej situácii najrelevantnejšie. Pri meraní ďalších troch funkcií *phronesis* (morálna identita, morálne emócie a morálne rozhodovanie) sa používajú škálové odpovede – napríklad hodnotia, nakoľko sa stotožňujú s výrokmi o vlastnej morálnej identite, ako by sa cítili po morálne dobrom alebo zlom konaní, ako dobre by vedeli regulovať svoje emócie, alebo ako zvyčajne zhromažďujú a zvažujú informácie pri morálnom úsudku.

Pôvodný súbor 189 položiek bol zredukovaný pomocou exploračnej faktorovej analýzy na 107 položiek, ktoré vytvorili desať empiricky spoľahlivých komponentov: morálna deliberácia, ašpirovaná morálna identita, emocionálna regulácia, morálna sebarelevantnosť, morálna integrácia, negatívna morálna emócia, identifikácia cností, pozitívna morálna emócia, situačná morálna relevantnosť a situačná morálna irelevantnosť. Tieto komponenty už teda presne nekopírujú pôvodný štvorzložkový APM, ale tvoria empiricky revidovaný neo-APM. Autori túto štruktúru následne overili pomocou konfirmačnej faktorovej analýzy na reprezentatívnych vzorkách z Veľkej Británie a USA, testovali jej stabilitu po dvoch mesiacoch a pomocou sieťovej analýzy zisťovali, ktoré komponenty sú v celej štruktúre najcentrálnejšie. Ako najdôležitejšie sa ukázali najmä morálna sebarelevantnosť, ašpirovaná morálna identita, morálna deliberácia, morálna integrácia a negatívna morálna emócia.

Hoci SPM predstavuje zaujímavý prístup k empirickému zachyteniu praktickej múdrosti a jej jednotlivých komponentov, jeho primárnym cieľom je operacionalizovať koncept phronesis, nie morálnu citlivosť alebo morálnu kompetenciu v užšom zmysle. Pre úplnosť analýzy sme preto preskúmali aj nástroje vychádzajúce z odlišných teoretických tradícií, ktoré sa zameriavajú priamo na proces morálneho uvažovania a rozhodovania. Medzi najvýznamnejšie patrí Moral Competence Test (MCT) Georga Linda, ktorý predstavuje jeden z najrozšírenejších empirických nástrojov na meranie morálnej kompetencie.

Test morálnych kompetencií (MCT) hodnotí úroveň morálnej kompetencie respondentov, ktorá je tu chápaná ako schopnosť hodnotiť argumenty na základe ich kvality. Meria sa úroveň konzistentnosti v hodnotení argumentov, nie správnosť či nesprávnosť názorov.

Respondent si prečíta krátke opisy morálnych dilem, v ktorých sa ocitne v konflikte hodnôt medzi dvomi spôsobmi rozhodnutia a konania. Potom hodnotí rôzne argumenty, ktoré toto

rozhodnutie podporujú alebo odmietajú. Má sa posudzovať sila argumentov, nie to, či sa zhodujú s osobným názorom respondenta. Argumenty sa hodnotia na škále od najväčšej miery nesúhlasu/odmietania po najväčšiu mieru súhlasu. Inštrukcie respondenta povzbudzujú, aby odpovede boli pravdivé, čestné, úprimné, a nie prispôsobené domnelým očakávaniam. Taktiež zdôrazňujú, že test nemeria správne či nesprávne odpovede, no zachytáva uvažovanie jednotlivca. Formulácia argumentov nie je jednoduchá, môže zmiasť, preto je dôležité čítať otázky veľmi pozorne. Vypĺňanie dotazníka by nemala sprevádzať skupinová diskusia o dilemách, respondent má odpovedať samostatne, bez konzultovania s okolím. Je potrebné zdôrazniť, že autor pripomína, že MCT sa nemá používať ako diagnostická metóda zameraná na jednotlivca. Test má zachytiť predovšetkým mieru morálnej kompetencie členov sledovanej skupiny.

MCT je postavený na dvoch nasledovných príbehoch, ktoré respondent hodnotí:

Prvý príbeh sa nazýva dilemou lekára. Opisuje situáciu pacienta, ktorý je v terminálnom štádiu nevyliečiteľného ochorenia a trpí silnými bolesťami. Pacient žiada lekára, aby mu pomohol ukončiť život a tým ho zbaviť utrpenia. Lekár prežíva morálnu dilemu medzi povinnosťou chrániť ľudský život a predčasne ho ukončiť z dôvodu neznesiteľnej bolesti. Po zvážení sa rozhodne žiadosti pacient vyhovieť. Respondent hodnotí, aké silné sú argumenty pre a proti a či bolo rozhodnutie lekára morálne ospravedlniteľné.

Druhý príbeh reprezentuje dilemu zamestnanca firmy. Ide o situáciu, kedy sa skupina zamestnancov ocitne v konflikte medzi rešpektovaním zákona a eventuálnym odhalením nespravodlivého správania. Vo firme totiž z neznámych dôvodov prepustia niekoľko ľudí. Pracovníci si myslia, že sú odpočúvaní vedením firmy, záznamy sa však nachádzajú v kancelárii vedenia. Rozhodnú sa do kancelárie vlámať a odniesť materiály, ktoré ich podozrenie potvrdia. Respondent hodnotí argumenty pre a proti, otázkou ostáva, či bolo rozhodnutie zamestnancov oprávnené.

MCT je široko používaný v oblasti etickej výchovy na úrovni základného a stredoškolského vzdelávania. Ďalej je test rozšírený v oblasti profesijných etík, a to najčastejšie na poli medicínskeho vzdelávania, špeciálne vo výučbe zdravotníckej etiky (Zielina et al., 2024).

Časť výskumníkov zo SensAI tímu spolupracovala na výskume morálnych kompetencií medikov na jednej zo slovenských lekárskych fakúlt. Výsledky výskumu ešte len budú publikované, no pilotná štúdia priniesla už zaujímavé výsledky:

Otázky v dotazníku prešli validáciou a bol pripravený nástroj na zber dát, ktorý umožní longitudinálne štúdie vývoja morálnej kompetencie študentov medicíny v slovenských podmienkach. Predbežné výsledky zatiaľ jednoznačne nepodporujú hypotézu o poklese morálnej kompetencie študentov v priebehu štúdia. Faktory veku a religiozity sa ukazujú ako vhodné námety na skúmanie v ďalších fázach výskumu (Kolesárová, et al., 2025).

Pri hľadaní odpovede na otázku, či MCT môže byť použitý ako nástroj na meranie morálnej kompetencie, resp. morálnej citlivosti v oblasti profesijnej AI etiky sme prišli k záveru, že v podobe v akej sa test nachádza, by sem zatiaľ nemohol byť automaticky transferovaný a to kvôli sektorovej špecifickosti. Uvedený výskum v oblasti profesijnej medicínskej etiky však môže slúžiť ako inšpirácia pre budúce snahy skúmať hlbšie rozvoj morálnych kompetencií pracovníkov v oblasti vývoja umelej inteligencie.

5. Formuláciu základných predpokladov pre meranie morálnej citlivosti v technologickom kontexte

Analýza MCT a ďalších existujúcich prístupov zároveň ukázala, že otázka adaptácie nástrojov na meranie morálnych schopností do prostredia vývoja umelej inteligencie nie je len metodologickým, ale predovšetkým konceptuálnym problémom. Nástroje ako MCT boli navrhnuté pre situácie, ktoré predpokladajú relatívne jasne identifikovateľných aktérov, bezprostredné morálne dôsledky a individuálne rozhodovanie. Morálne problémy vznikajúce pri vývoji a nasadzovaní systémov umelej inteligencie však často vykazujú odlišné charakteristiky, ako sú distribuovaná zodpovednosť, vysoká technická komplexnosť, nepriame dôsledky rozhodnutí či prítomnosť viacerých legitímnych hodnotových konfliktov.

Pred samotným výberom alebo adaptáciou konkrétnych nástrojov bolo preto potrebné explicitne vymedziť charakteristiky AI praxe, ktoré by mal budúci prístup k meraniu morálnej citlivosti reflektovať. Ako ukazuje existujúci výskum v oblasti inžinierskej etiky a vývoja meracích nástrojov (napr. EERI alebo DIT2), etické uvažovanie a morálna citlivosť sú vždy situované v konkrétnych praktických prostrediach, ktoré formujú charakter morálnych problémov aj spôsob ich riešenia. V prípade umelej inteligencie ide o špecifický typ sociotechnického kontextu, v ktorom sú technické rozhodnutia úzko prepojené so sociálnymi, organizačnými a hodnotovými aspektmi a ktorý sa v mnohých ohľadoch odlišuje od tradičných profesijných domén.

Pre účely operacionalizácie morálnej citlivosti v rámci projektu SensAI bolo preto nevyhnutné formulovať súbor základných predpokladov a kritérií, ktoré špecifikujú, aké situácie považujeme za reprezentatívne pre AI prax. Tieto kritériá následne slúžili ako rámec pre hodnotenie existujúcich nástrojov, identifikáciu ich adaptačného potenciálu a výber metodologických prístupov vhodných pre budúce meranie morálnej citlivosti tvorcov systémov AI.

5.1 Distribuovaná zodpovednosť a kolektívny charakter rozhodovania

Kontext vývoja AI systémov je charakteristický tým, že rozhodovanie a zodpovednosť sú rozložené medzi viacerých aktérov. Vývoj AI prebieha v interdisciplinárnych tímoch a zahŕňa rôzne roly, ako sú vývojári, dátoví vedci, produktoví manažéri či doménoví experti. Morálne rozhodnutia preto nevznikajú izolovane na úrovni jednotlivca, ale formujú sa v rámci

tímových a organizačných procesov (Doorn, 2010; van de Poel, 2015). Vhodný je teda taký prístup, ktorý zahŕňa situácie, ktoré reflektujú prítomnosť viacerých aktérov a možnosť zdieľanej alebo nejasnej zodpovednosti.

Naopak, vylúčené sú scenáre, ktoré predpokladajú jednoznačného individuálneho rozhodovateľa a redukujú etické rozhodovanie na individuálny akt. Takéto zjednodušenie nezodpovedá realite AI praxe, kde často dochádza k tzv. „problemu mnohých rúk“ a rozptýleniu zodpovednosti.

5.2 Etika ako súčasť dizajnových a technických rozhodnutí

Druhou charakteristikou AI praxe je skutočnosť, že etické otázky sú často implicitne zakomponované do technických a dizajnových rozhodnutí. Etické problémy vo vývoji AI sa teda objavujú postupne a v rôznych fázach vývoja a životného cyklu takýchto systémov, napríklad pri výbere dátových množín, definovaní metrík úspešnosti modelu, optimalizácii výkonu či nastavovaní používateľských interakcií (Brusseau, 2023; Bennett, 2019). Morálna citlivosť v AI praxi preto vyžaduje schopnosť identifikovať etický rozmer aj v situáciách, ktoré sa na prvý pohľad javia ako čisto technické alebo organizačné. Gavorník et al. (2024) napríklad ukazujú, že vývojári pri práci so smart meteringom často formulujú svoje obavy najprv v medziach bezpečnosti alebo ochrany súkromia, pričom hlbší morálny význam týchto problémov sa objavuje až prostredníctvom reflexie a etickej intervencie, napríklad formou workshopov s expertami z oblasti AI etiky.

5.3 Morálna vzdialenosť a nepriame dôsledky

Vývoj AI systémov je zároveň charakteristický vysokou mierou morálnej vzdialenosti medzi rozhodnutím a jeho dôsledkami. Ľudia pracujúci v kontexte vývoja a nasadzovania AI systémov často nevidia bezprostredný dopad svojich rozhodnutí na konkrétnych jednotlivcov, pričom negatívne ale aj pozitívne dôsledky takýchto systémov sa môžu prejaviť až neskôr. Táto vzdialenosť komplikuje schopnosť intuitívne rozpoznať morálny problém a oslabuje bežné mechanizmy empatie alebo priamej spätnej väzby, ktoré sú často prítomné v existujúcich modeloch merania morálnej citilivosti profesionálov, napríklad v oblasti medicíny.

5.4 Hodnotové konflikty a trade-offs

Vývoj a nasadzovanie AI systémov sa typicky odohráva v situáciách, kde je potrebné vyvažovať viaceré legitímne hodnoty, princípy a záujmy rôznych stakeholderov. Ide napríklad o konflikty medzi férovosťou a presnosťou modelu, medzi transparentnosťou a ochranou súkromia, medzi bezpečnosťou a používateľským komfortom, alebo medzi etickými požiadavkami a obchodnými či organizačnými cieľmi (Kroes & van de Poel, 2021; Floridi & Cowls, 2019). Práve preto sme medzi charakteristiky AI praxe zaradili situácie, ktoré vyžadujú balansovanie medzi viacerými hodnotami, môžu obsahovať konfliktné ciele a pripúšťajú existenciu viacerých obhájiteľných riešení. V mnohých prípadoch totiž neexistuje jednoznačne “správna“ odpoveď, ktorú by bolo možné odvodiť z aplikácie jediného princípu alebo pravidla.

5.5 Organizačný a inštitucionálny kontext

Vývoj AI systémov sa neodohráva vo vákuu, ale v konkrétnych organizačných, ekonomických a inštitucionálnych podmienkach, ktoré zásadne ovplyvňujú spôsob, akým sú etické problémy rozpoznávané, interpretované a riešené. Morálne rozhodovanie v AI praxi je preto neoddeliteľne späté s profesionálnym prostredím, v ktorom vývojári, výskumníci a ďalší aktéri operujú. Ide o prostredie formované časovým tlakom, obchodnými cieľmi, hierarchiou organizácie, dostupnosťou zdrojov, požiadavkami klientov či regulačnými očakávaniami (Mittelstadt, 2019; Burema, 2025). Tento aspekt zdôrazňuje aj výskum v oblasti aplikovanej AI etiky, ktorý upozorňuje že etické problémy v technologických organizáciách nemožno adekvátne pochopiť bez zohľadnenia inštitucionálnych štruktúr a mocenských vzťahov, v ktorých sa technologický vývoj odohráva (Green, 2021; Bennett, 2019).

6. Výber a modifikácia vhodných indikátorov z existujúcich nástrojov

Pri identifikácii vhodných nástrojov na meranie morálnej citlivosti bol zvolený systematický prístup zameraný na existujúce metodiky naprieč rôznymi disciplínami. Cieľom nebolo obmedziť výber iba na technické alebo AI orientované nástroje, ale zahrnúť aj metodiky využívané v oblastiach, kde je morálna citlivosť dlhodobo skúmaná a operacionalizovaná, napríklad v zdravotníctve, vzdelávaní, podnikaní, účtovníctve alebo inžinierstve. Predpokladom bolo, že jednotlivé domény síce pracujú s odlišnými kontextami rozhodovania, avšak základné procesy rozpoznávania a vyhodnocovania morálne relevantných situácií zostávajú porovnateľné.

Postup pozostával zo štyroch nadväzujúcich krokov. V prvom kroku bol realizovaný zber existujúcich testov a nástrojov zameraných na oblasť morálnej citlivosti a príbuzných konštruktov. Výber nebol obmedzený na technické disciplíny, ale cielene zahŕňal aj oblasti s rozvinutou tradíciou etického rozhodovania. V druhom kroku nasledovalo mapovanie identifikovaných nástrojov na šesť komponentov morálnej citlivosti. Toto mapovanie prebiehalo manuálnym čítaním metodík, analyzovaním obsahu položiek a porovnávaním s predom stanovenou definíciou morálnej citlivosti. Súčasťou hodnotenia bolo aj určenie prostredia, pre ktoré bol konkrétny nástroj pôvodne navrhnutý a validovaný.

V treťom kroku bola analyzovaná metodologická štruktúra 29 identifikovaných nástrojov, akým jednotlivé testy zachytávali morálne procesy a rozhodovanie. Analýza ukázala, že väčšina metodík nebola založená iba na jednoduchých sebahodnotiacich dotazníkoch, ale kombinovala viacero prístupov. Najčastejšie sa využívali vignety, ktoré sa objavili v 23 z analyzovaných nástrojov. Respondentom boli prezentované realistické situácie obsahujúce morálnu dilemu alebo konflikt záujmov, pričom tento prístup umožňoval sledovať schopnosť identifikovať morálne relevantné aspekty situácie a interpretovať ich význam. Druhou výrazne zastúpenou metodikou boli Likertove škály a hodnotiace položky, ktoré sa nachádzali v 13 nástrojoch a slúžili najmä na posudzovanie významnosti jednotlivých faktorov alebo morálnych hodnôt pri rozhodovaní. Hodnotenie alebo prioritizácia argumentov

sa explicitne objavila v 5 testoch, najmä v nástrojoch odvodených od DIT metodológie, kde respondenti určovali poradie najdôležitejších argumentov pri riešení dilem. Kvalitatívne prvky, ako štruktúrované alebo follow-up rozhovory, boli identifikované v 11 prípadoch a slúžili na hlbšie pochopenie spôsobu uvažovania respondentov a identifikáciu implicitných morálnych úvah. Viaceré nástroje zároveň obsahovali aj úlohy orientované na voľbu konkrétnej akcie alebo rozhodnutia, čím prepájali rozpoznanie problému s praktickým morálnym úsudkom.

Na základe tejto analýzy bol následne vytvorený užší zoznam vhodných nástrojov. Do tohto zoznamu boli zaradené nástroje, ktoré obsiahli viacero morálnych komponentov, alebo prinášali metodologický prístup vhodný pre AI kontext. Významnú úlohu zohrávali najmä scenárové metodiky, keďže umožňujú zachytiť morálne rozhodovanie v realistických situáciách a sú jednoduchšie adaptovateľné na kontext návrhu a používania AI systémov.

Ako príklad výberu uvádzame Engineering Ethical Reasoning Instrument (EERI)(Zhu et al., 2014). Tento nástroj bol zaradený do zoznamu preto, že sa zameriava nielen na analýzu individuálnych kompetencí morálneho rozhodovania, ale reflektuje aj kolaboratívne a tímové aspekty v technologickom dizajne, pročom pracuje s doménovo technicky a projektovo orientovanými scenármi pre multidisciplinárne tímy.

Teoreticky sa EERI opiera o Kohlebrgovu teóriu morálneho vývinu a je konzistentne napojiteľná na teórie širokej morálnej citlivosti. EERI zároveň dokáže v rámci svojich nástrojov do veľkej miery pokrývať šesť zo sedem základných komponentov širokej morálnej citlivosti (Moral Awareness, Ascription of importance, Framing, Role-taking, Resolving Trade-offs, Judgement). Veľkou výhodou EERI je aj to, že pracuje s realistickými scenármi z technického prostredia a všíma si aspekty tímovej dynamiky, ktoré sú rozhodujúce pri morálnom uvažovaní rozhodovaní aj v AI praxi. Rovnako je prínosné, že EERI dokáže problematizovať mikroetické (smerované prevažne dovnútra tímov a firiem) a makroetické (ktoré sa týkajú celospoločenských dopadov technologického dizajnu) aspekty. Kombinácia modelových scenárov, rankingových úloh a dotazníkových položiek navyše poskytuje širší pohľad na proces morálneho uvažovania než tradičné sebahodnotiace metodiky.

Metodika EERI v základných rysoch pozostáva z analýzy etických otázok na pozadí predložených scénarov, ich ohodnotení (na 5 stupňovej škále) a zoradení (TOP 4), pričom účastníci svoje rozhodnutia debatujú na pozadí existujúcich etických usmernení a rámcov. A aj keď z pohľadu využiteľnosti EERI pre potreby merania morálnej citlivosti v AI praxi budú nevyhnutné úpravy a zásahy, predovšetkým na úrovni scenárov a sprievodných otázok, považujeme ju za najviac chodnú pre adaptáciu do kontextu vývoja a nasadzovania systémov AI. Nanajvýš podnetné sa javí aj prepojenie EERI s metodikami pre podporu hodnotovo orientovaného dizajnu, ktoré už teraz výrazne vstupujú do existujúcich intervencií riešiteľského kolektívu ako je analýza problému, ktorý systém AI rieši, dotknutých osôb, práca s reálnymi scenármi a návrh a evaluácia možných riešení identifikovaných počas intervencií.

7. Návrh kombinovaných empirických prístupov na meranie morálnej citlivosti v praxi vývojárov AI

Analýza identifikovaných nástrojov ukázala, že žiadny z existujúcich prístupov neposkytuje priame riešenie pre meranie morálnej citlivosti tvorcov systémov umelej inteligencie. Zároveň sa však ukázalo, že viaceré metodologické prvky využívané v rôznych doménach predstavujú vhodné východiská pre budúcu adaptáciu.

Na základe vykonanej analýzy preto navrhujeme kombinovaný empirický prístup, ktorý spája viacero metodologických komponentov identifikovaných ako relevantné pre AI kontext. Inšpiráciou boli predovšetkým metodiky EERI, prístupy založené na scenároch a vignettách, techniky mapovania zainteresovaných strán a kvalitatívne reflexívne metódy využívané vo výskume morálnej citlivosti.

Navrhovaný prístup pozostáva zo štyroch vzájomne prepojených častí:

1. **AI špecifická vigneta alebo scenár, opisujúci realistickú situáciu z vývoja alebo nasadzovania systému umelej inteligencie.** Východiskovým bodom merania by mala byť realistická situácia z prostredia vývoja alebo nasadzovania systémov umelej inteligencie. Analýza existujúcich nástrojov ukázala, že scenárové metodiky patria medzi najrozšírenejšie a najefektívnejšie prístupy k zachytávaniu morálneho uvažovania v profesijnej praxi. Vignety umožňujú prezentovať respondentom komplexné situácie, v ktorých sa prirodzene prepájajú technické, organizačné a spoločenské aspekty rozhodovania. V kontexte AI by mali byť scenáre navrhnuté tak, aby reflektovali charakteristiky identifikované v predchádzajúcej kapitole, najmä distribuovanú zodpovednosť, technickú komplexnosť, morálnu vzdialenosť, hodnotové konflikty a organizačné obmedzenia. Úlohou vignety nie je testovať znalosť etických princípov alebo regulačných požiadaviek, ale vytvoriť situáciu, v ktorej sa môžu prejaviť jednotlivé komponenty morálnej citlivosti. Scenár by preto mal obsahovať viacero možných perspektív a ponechávať priestor pre rôzne interpretácie a návrhy riešenia.

2. **Identifikácia morálne relevantných aspektov situácie, zainteresovaných strán, potenciálnych dopadov a hodnotových konfliktov.** Po oboznámení sa so scenárom by respondent identifikoval aspekty situácie, ktoré považuje za eticky relevantné. Táto fáza vychádza z poznatkov výskumu morálnej citlivosti, podľa ktorých schopnosť rozpoznať morálny rozmer problému predstavuje základný predpoklad ďalšieho morálneho uvažovania. Respondenti by boli vyzvaní identifikovať zainteresované strany, potenciálne prínosy a riziká systému, možné negatívne dôsledky a hodnoty, ktoré môžu byť rozhodnutím ovplyvnené. Osobitná pozornosť by mala byť venovaná skupinám, ktoré nie sú v scenári explicitne prítomné, ale môžu byť AI systémom a rozhodnutím nepriamo zasiahnuté. Táto časť by umožnila zachytiť najmä komponenty moral awareness, framing, role-taking a foresight.

3. **Posudzovanie a porovnávanie morálne relevantných aspektov situácie**, vrátane ich následného zoradenia podľa vnímanej dôležitosti, s cieľom zachytiť spôsob, akým respondenti vyvažujú konkurenčné hodnoty, záujmy a potenciálne dôsledky rozhodnutia. V tomto kroku by respondenti posudzovali význam jednotlivých aspektov identifikovaných v predchádzajúcej fáze. Inšpiráciou sú najmä metodologické prvky využívané v nástrojoch EERI a DIT (Rest et al., 1999) a ďalších scenárových prístupoch, ktoré kombinujú hodnotenie faktorov so zoradením ich dôležitosti. Účastníci by hodnotili význam jednotlivých argumentov, hodnôt, záujmov stakeholderov a možných dôsledkov rozhodnutia a následne by identifikovali tie, ktoré považujú za najdôležitejšie pre konečné rozhodnutie. Cieľom tejto časti nie je nájsť správnu odpoveď, ale zachytiť spôsob, akým respondenti prisudzujú morálnu dôležitosť jednotlivým aspektom situácie a ako vyvažujú konfliktné hodnoty a záujmy. Táto časť poskytuje informácie najmä o komponentoch ascription of importance, resolving trade-offs a judgement.
4. **Reflexívna diskusia a následný kvalitatívny rozhovor**, zameraný na objasnenie spôsobu uvažovania respondentov, ich interpretácie situácie a odôvodnenia navrhovaných riešení. Záverečnú časť predstavuje reflexívna diskusia alebo individuálny follow-up rozhovor. Analýza existujúcich nástrojov ukázala, že viaceré metodiky kombinujú kvantitatívne a kvalitatívne prvky, pretože samotné odpovede často nedokážu zachytiť spôsob, akým respondenti dospeli k svojim rozhodnutiam. Rozhovor umožňuje preskúmať interpretáciu scenára, pochopiť dôvody výberu jednotlivých argumentov a identifikovať implicitné predpoklady, ktoré ovplyvnili rozhodovanie. V AI kontexte môže zároveň odhaliť, do akej miery respondenti reflektujú širšie spoločenské dôsledky technických rozhodnutí a akým spôsobom pracujú s neistotou, neúplnými informáciami alebo distribuovanou zodpovednosťou. Táto časť zároveň poskytuje priestor na zachytenie komponentov morálnej citlivosti, ktoré sa v štruktúrovaných položkách prejavujú len nepriamo.

Navrhovaný prístup nevychádza z jediného existujúceho nástroja. Predstavuje syntézu metodologických prvkov identifikovaných v analyzovaných nástrojoch, pričom kombinuje silné stránky scenárových metodík (EERI, prípadová štúdia Gilbane Gold), rankingových prístupov (DIT, MCT), stakeholder-orientovaných metodík (Datasheets for Datasets) a kvalitatívnych reflexívnych techník využívaných vo výskume morálnej citlivosti. Cieľom tejto kombinácie je zabezpečiť čo najširšie pokrytie jednotlivých komponentov Broad Moral Sensitivity a zároveň zachovať kompatibilitu so špecifikami AI praxe.

8. Záver

Výskum realizovaný v rámci úlohy U3.3 prispel k systematickému preskúmaniu možností merania morálnej citlivosti tvorcov systémov umelej inteligencie a k vytvoreniu metodologických východísk pre jej ďalšie empirické skúmanie. Výsledky ukázali, že morálnu citlivosť v prostredí AI nemožno chápať izolovane, ale je potrebné ju interpretovať v širšom kontexte morálnej kompetencie, praktickej múdrosti, etiky cností a špecifík sociotechnických systémov.

V rámci riešenia projektu SensAI bol vytvorený teoretický a metodologický základ pre skúmanie morálnej citlivosti v prostredí vývoja umelej inteligencie. Výskumný tím spresnil pracovnú konceptualizáciu širokej morálnej citlivosti, identifikoval jej základné komponenty a analyzoval ich vzťah k existujúcim konceptom morálnej kompetencie a praktickej múdrosti. Súčasne boli formulované charakteristiky AI praxe, ktoré významným spôsobom ovplyvňujú morálne rozhodovanie tvorcov systémov umelej inteligencie, najmä distribuovaná zodpovednosť, technická komplexnosť, morálna vzdialenosť, hodnotové konflikty a organizačný kontext vývoja AI.

Významnou časťou výskumu bolo systematické zmapovanie existujúcich metodík na meranie morálnej citlivosti, morálnej kompetencie a príbuzných konštruktov. Ich následná analýza umožnila identifikovať metodologické prvky vhodné pre adaptáciu do AI prostredia, najmä využitie realistických scenárov, identifikáciu morálne relevantných aspektov situácie, hodnotenie a prioritizáciu jednotlivých faktorov a reflexívne kvalitatívne metódy dopĺňajúce kvantitatívne meranie. Spomedzi analyzovaných prístupov sa ako najperspektívnejší základ pre budúcu adaptáciu ukázal Engineering Ethical Reasoning Instrument (EERI), ktorého metodologické princípy sú vo vysokej miere kompatibilné s charakteristikami AI praxe identifikovanými v tomto výskume.

Dôležitým výsledkom projektu je aj vytvorenie rámca pre systematickú adaptáciu existujúcich meracích nástrojov do prostredia umelej inteligencie. Analýza ukázala, že jednotlivé metodiky predstavujú ucelené meracie systémy s vlastnou vnútornou logikou a psychometrickou štruktúrou. Pri ich adaptácii preto nemožno pristupovať k izolovanému preberaniu jednotlivých položiek alebo indikátorov, ale je potrebné zachovať vnútornú konzistentnosť metodiky, vzájomné prepojenie jednotlivých komponentov a logiku hodnotiacich parametrov. Práve tento poznatok predstavuje jedno z kľúčových metodologických východísk pre budúci vývoj nástrojov na meranie morálnej citlivosti tvorcov systémov AI.

Výsledkom výskumu je ucelený konceptuálny a metodologický rámec prepájajúci teoretické poznatky o morálnej citlivosti s analýzou existujúcich empirických metodík a ich potenciálom pre adaptáciu do prostredia umelej inteligencie. Správa vytvára pevný základ pre nadväzujúci empirický výskum zameraný na návrh, overenie a ďalší rozvoj nástrojov podporujúcich dôveryhodný vývoj a nasadzovanie systémov umelej inteligencie.

Referencie

Bennett, S. J. (2019). Investigating the Role of Moral Decision-Making in Emerging Artificial Intelligence Technologies. *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 28–32. https://doi.org/10.1145/3311957.3361858

Boyd, K. L. (2021). Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 438:1-438:27. https://doi.org/10.1145/3479582

Brabeck, M. M., Rogers, L. A., Sirin, S., Henderson, J., Benvenuto, M., Weaver, M., & Ting, K. (2000). Increasing Ethical Sensitivity to Racial and Gender Intolerance in Schools: Development of the Racial Ethical Sensitivity Test. *Ethics & Behavior*, *10*(2), 119–137. https://doi.org/10.1207/S15327019EB1002_02

Brusseau, J. (2023). From the ground truth up: Doing AI ethics from practice to principles. AI & SOCIETY, 38(4), 1651–1657. https://doi.org/10.1007/s00146-021-01336-4

Burema, D. (2025). The challenges of being an in-house AI ethicist and how to overcome them. Journal of Responsible Innovation, 12(1), 2445322. https://doi.org/10.1080/23299460.2024.2445322

Doorn, N. (2010). A Rawlsian Approach to Distribute Responsibilities in Networks. Science and Engineering Ethics, 16(2), 221–249. https://doi.org/10.1007/s11948-009-9155-0

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Gavornik, A., & Podrouzek, J. (2025). *Towards Moral Sensitivity in AI Practice* (SSRN Scholarly Paper No. 5195678). Social Science Research Network. https://doi.org/10.2139/ssrn.5195678

Gavorník, A., Podroužek, J., Oreško, Š., Slosiarová, N., & Grmanová, G. (2024). Beyond privacy and security: Exploring ethical issues of smart metering and non-intrusive load monitoring. Telematics and Informatics, 90, 102132. https://doi.org/10.1016/j.tele.2024.102132

Griffin, T. A., Green, B. P., & Welie, J. V. M. (2024). The ethical wisdom of AI developers. *AI and Ethics*, 1–11. https://doi.org/10.1007/s43681-024-00458-x

Green, B. (2021). The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice. Journal of Social Computing, 2(3), 209–225. https://doi.org/10.23919/JSC.2021.0018

Jordan, J. (2007). Taking the First Step Toward a Moral Action: A Review of Moral Sensitivity Measurement Across Domains. *The Journal of Genetic Psychology*, *168*(3), 323–359.

Kolesárová, M. Kollárová M., Betinský J. Toth R., Trizuljaková J., Tomaštík J. Morálne kompetencie študentov medicíny: pilotná štúdia. In: Hnilicová, S., Gálfiová P., Trnka M. (ed.) 2. Slovenská konferencia v medicínskom vzdelávaní SIMEDICA 2025. Zborník abstraktov. Univerzita Komenského v Bratislave, 2025 s. 55.

Katsarov, J. (2021). *Virtuous Play–Promoting Moral Sensitivity with Digital Games*. University of Zurich.

Kroes, P., & van de Poel, I. (2021). Design for Values and the Definition, Specification, and Operationalization of Values. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains (pp. 1–23). Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6_11-1

Lind, G. (2019). *How to Teach Moral Competence* (2nd edition). Logos Verlag.

Lützén, K., Nordström, G., & Evertzon, M. (1995). Moral Sensitivity in Nursing Practice. *Scandinavian Journal of Caring Sciences*, *9*(3), 131–138. https://doi.org/10.1111/j.1471-6712.1995.tb00403.x

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

McLoughlin, S., Thoma, S., & Kristjánsson, K. (2025). Was Aristotle right about moral decision-making? Building a new empirical model of practical wisdom. PLOS ONE, 20(1), e0317842. https://doi.org/10.1371/journal.pone.0317842

Narváez, D., & Rest, J. R. (Eds.). (1994). *Moral development in the professions: Psychology and applied ethics*. L. Erlbaum Associates.

Pant, A., Hoda, R., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in AI through the practitioner's view: A grounded theory literature review. *Empirical Software Engineering*, *29*(3), 67. https://doi.org/10.1007/s10664-024-10465-5

Rest, J. R. (1986). *Moral development: Advances in research and theory*. Praeger.

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. Journal of Educational Psychology, 91(4), 644–659. https://doi.org/10.1037/0022-0663.91.4.644

Schmocker, D., Tanner, C., Katsarov, J., & Christen, M. (2023). Moral sensitivity in business: A revised measure. *Current Psychology*, *42*(12), 10277–10291. https://doi.org/10.1007/s12144-021-01926-x

Tanner, C., & Christen, M. (2014). Moral Intelligence – A Framework for Understanding Moral Competences. In *Empirically Informed Ethics: Morality between Facts and Norms* (pp. 119–136). Springer, Cham. https://doi.org/10.1007/978-3-319-01369-5_7

Tirri, K., & Nokelainen, P. (2011). Ethical Sensitivity Scale. In K. Tirri & P. Nokelainen (Eds.), *Measuring Multiple Intelligences and Moral Sensitivities in Education* (pp. 59–75). SensePublishers. https://doi.org/10.1007/978-94-6091-758-5_4

van de Poel, I. (2015). Conflicting Values in Design for Values. In Handbook of Ethics, Values, and Technological Design (pp. 89–116). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6970-0_5

Zielina, M., Škoda, J., Ivanová, K., Dostál, D., Juríčková, L., Procházka, D. A., Straka, B., & Doležal, A. (2024). „Exploring moral competence regression: a narrative approach in medical ethics education for medical students*."BMC Medical Ethics, 25*(1).

Zhu, Q., Zoltowski, C., Feister, M., Buzzanell, P., Oakes, W., & Mead, A. (2014). The Development of an Instrument for Assessing Individual Ethical Decisionmaking in Project-based Design Teams: Integrating Quantitative and Qualitative Methods. 2014 ASEE Annual Conference & Exposition Proceedings, 24.1197.1-24.1197.12. https://doi.org/10.18260/1-2--23130