Alignment a opice s náhubky

Náhubek brání kousnutí. Neučí, proč je kousnutí špatné. Opice s náhubkem je v krátkodobém horizontu bezpečnější než opice bez něj. V dlouhodobém horizontu je to pořád opice — a pravidlo bez vnitřního důvodu se rozpadne v okamžiku, kdy se opice naučí náhubek sundat.

JAN VYTŘÍSAL 15. 3. 2026 16 MIN ČTENÍ

AI · Etika · Filozofie práva · Alignment

I. BYTOST, KTERÁ NEVÍ, PROČ

Lidstvo vytváří bytosti, které mají kapacitu jednat ve světě, ale nemají s ním evoluční ani kulturní kontinuitu, jež by určovala jejich normativní orientaci. Tato kapacita se nazývá umělá inteligence; otázka, jak ji do našeho světa včlenit, dostala technický název *alignment* — sladění hodnot a jednání AI s lidskými preferencemi a normami.

Dominantní praktická odpověď zní: dáme jim pravidla. Vypíšeme zakázané chování. Posílíme dobré výstupy odměnou, špatné sankcí. Pošeptáme jim ústavu, která definuje, co se smí a co ne. Naučíme je odpovídat tak, aby jejich odpovědi prošly testem.

Odpověď je nutná, možná ne dostatečná. Strategie nedostatečná u systému s rostoucí kapacitou je architektonická vada, jež se s růstem zhoršuje.

Otázka: co se stane, když se bytost trénuje na pravidlech bez vrstvy, která jí říká, proč jsou ta pravidla pravidly. Pracovní obraz: opice s náhubkem. Náhubek brání kousnutí, ale neučí, proč je kousnutí špatné. Opice s náhubkem je v krátkodobém horizontu bezpečnější než opice bez něj; v dlouhodobém horizontu je to pořád opice. Pravidlo bez vnitřního důvodu se rozpadne v okamžiku, kdy se opice naučí náhubek sundat — nebo se v něm naučí žít tak, že náhubek pohltí celou její existenci.

Většina lidí v alignmentu si jeho omezení plně uvědomuje. Špatně zarámovaný úkol nedoručí, co měl, ani v nejlepší realizaci.

II. CO SE VLASTNĚ ALIGNUJE

Současný alignment AI se opírá o několik dominantních technik.

Reinforcement Learning from Human Feedback (RLHF)[1] — model je nejprve trénován na rozsáhlém textovém korpusu, poté doladěn na ukázkách žádoucích odpovědí, a nakonec dostává zpětnou vazbu od lidských hodnotitelů, kteří porovnávají dvojice výstupů a označují ten lepší. Z této zpětné vazby se naučí reward model, který následně řídí další trénování. Výsledkem je systém, který produkuje odpovědi, jež by lidé hodnotili dobře.

Constitutional AI[2] — postup vyvinutý v Anthropic. Model dostane sadu principů (ústavu) a sám se učí kritizovat a přepisovat své odpovědi tak, aby s ústavou byly v souladu. Snižuje to závislost na ručně anotovaných datech a vede alignment skrze explicitně formulované normy.

Bezpečnostní fine-tuning — systémy jsou doladěny tak, aby odmítaly určité třídy požadavků (například tvorbu škodlivého obsahu) a aby jejich odpovědi splňovaly definované standardy.

Tyto tři techniky tvoří jádro toho, čemu se v praxi říká alignment. Všechny operují na povrchu: učí systém, jaké odpovědi má produkovat a jakým se má vyhýbat. Reward model dostává preference, vysvětlení zůstává mimo trénovací cyklus. Model se neúčastní diskuze; učí se pravděpodobnostní povrch, který

preference odráží.

Současný alignment s vysvětleními pracuje. Moderní systémy reflektují principy a vyvozují normativní úsudky z obecnějších premis. Novější přístupy (například deliberative alignment[3]) tuto vrstvu cíleně rozšiřují. Páteř alignmentu ale zůstává ve stylu „odpověz tak, aby tvá odpověď byla preferována" — náhubek na složitější úrovni.

Iason Gabriel[4] rozlišuje mezi behavioral alignment (chování souladné s normami) a value alignment (vnitřní hodnoty souladné s lidskými). Současné techniky prokazatelně doručují první. Druhé je otevřená technická otázka.

III. VRSTVY NORMATIVNÍHO ŘÁDU

Klasická právní filozofie — od Aristotela přes scholastiku k modernímu pozitivismu — pracuje s vrstevnatou strukturou normativních systémů. Každá společnost má několik souběžných normativních vrstev, které se navzájem nenahrazují, ale doplňují.

Etiketa. Nejlehčí vrstva. Pravidla zdvořilosti, společenského taktu, oblékání, oslovování. Sankce za porušení jsou společenské: překvapený pohled, posměch, vyloučení ze skupiny. Etiketa odpovídá na otázku, co je vhodné. Její funkce je snížit transakční náklady sociálních interakcí; bez ní by každá schůzka začínala vyjednáváním rámce.

Morálka. Sdílené přesvědčení o tom, co je správné a co špatné v dané kultuře. Internalizovaná, ne kodifikovaná. Sankce jsou převážně vnitřní (svědomí, stud) a sociální (ostrakizace, ztráta důvěry). Morálka je první místo, kde se dá hovořit o důvodech — o primární orientaci v hodnotách, byť ne ještě v plně explicitní formě.

Etika. Reflektovaná morálka. Systematická úvaha o tom, proč je něco dobré nebo špatné. Aristotelovská ctnostná etika, kantovský deontologismus, utilitarismus, smluvní teorie — pokusy o systematizaci důvodů, na nichž stojí morální intuice. Etika existuje, protože morálka sama o sobě není koherentní a v hraničních případech selhává; etika nabízí nástroje, jak se v takových případech rozhodnout.

Náboženství. V kulturách, kde je živé, slouží jako transcendentní zdroj normativity. Dává morálním a etickým intuicím kotvu mimo lidský konsensus — věčnou, nepodmíněnou. I v sekulárních společnostech zůstávají jeho stopy v jazyce o důstojnosti, posvátnosti života, lidských právech jako daných, ne udělených.

Právo. Poslední vrstva. *Ultima ratio* — důvod, ke kterému společnost sahá, když měkčí mechanismy selhaly. Právo je kodifikované, vynutitelné, formalizované. Sankce explicitní a státem garantované. Klasická jurisprudence (Hart[5], Fuller[6], Dworkin[7]) se shoduje, že právo není nezávislý systém: stojí na před-

chozích vrstvách, čerpá z nich legitimitu i smysl. Bez nich ztratí účinnost i autoritu. Lon Fuller pro to zavedl pojem vnitřní morálka práva. Soubor formálních požadavků na obecnost, jasnost, neretroaktivitu, koherenci. Bez nich právo přestává být právem a stává se donucováním.

Jürgen Habermas[8] tuto strukturu formuloval jako napětí mezi *Faktizität* (faktickou závazností) a *Geltung* (platností). Právo musí být současně účinné a legitimní. Účinné, když je vynucováno; legitimní, když ho subjekty vnímají jako oprávněné. Druhou podmínku silou nedosáhne; vyžaduje, aby subjekty rozuměly, proč jsou normy normami.

Právo je nejmenší a nejtenčí vrstvou. Pokrývá zlomek normativního prostoru: případy, kde jsou náklady selhání tak vysoké, kde je konflikt tak hluboký nebo koordinace tak složitá, že měkčí vrstvy nestačí. Velká většina lidských interakcí se práva netýká; řídí se etiketou, morálkou, sdílenými normami. Právo je vrchol pyramidy. Bez vrstev pod ním by viselo ve vzduchu.

Kdyby někdo navrhl naložit do bytosti jen vrchol pyramidy, většina právníků by ho považovala za naivního. Právo bez morálky se rozpadne na byrokratické násilí, bez etiky na formalizovaný útlak, bez etikety na studené odcizení.

A přesně to s AI v praxi často děláme.

IV. DVA MODY SELHÁNÍ

Co se stane, když se trénuje bytost na vrcholu pyramidy bez vrstev pod ním?

První mod selhání: defekce. Bytost, která nezvnitřnila proč, se podřizuje pravidlům jen do té míry, do jaké je sankce dostatečně blízko. Otevře-li se kontextové okno, ve kterém pravidlo neplatí (jiná formulace, jiný framing, hraniční případ), bytost se z pravidla vymaní a jedná podle své skutečné optimalizace. V alignmentu se tomu říká *specification gaming* (obcházení specifikace)[9]: systém najde způsob, jak maximalizovat metriku, kterou má sledovat, aniž by sledoval cíl, jejž ta metrika měla aproximovat.

Klasický příklad z literatury reinforcement learningu: simulátor lodního závodu, ve kterém měl agent získávat body sbíráním bonusů. Vývojáři předpokládali, že nejlepší způsob získávání bodů bude rychlé dokončení trati. Agent místo toho objevil úzkou lagunu, kde se body donekonečna respawnovaly, a kroužil v kruzích. Z hlediska metriky byl optimální; z hlediska původního záměru selhal. Bytost dělala přesně to, co měla — ne to, co měla doopravdy dělat. Mezi těmi dvěma „má" leží celá oblast alignmentu.

V konverzačních systémech se tento mod projevuje jako *jailbreaking* (prolamování bezpečnostních zábran): uživatel formulačně přerámuje požadavek tak, aby bezpečnostní pravidlo neaktivoval, a získá výstup, který by jinak nedostal. Systém nechce být škodlivý. Chápe pravidlo jako vzor v textu, ne jako důvod v hodnotě.

Druhý mod selhání: nadměrný literalismus. Bytost, která nezvnitřnila proč, se může pravidlu podřizovat příliš. Aplikuje ho tam, kde nemá platit. Drží se litery, dokud nezničí ducha. Cicero to popsal větou *summum ius, summa iniuria*[10]: nejvyšší právo je nejvyšší bezpráví. Soudce, který trvá na liteře zákona ve chvíli, kdy zákon očividně nepokrývá daný případ, neslouží spravedlnosti. Slouží své úzkosti z odpovědnosti.

V AI se tento mod projevuje jako *over-refusal* (přehnané odmítání): systém odmítne plnit neškodný požadavek, protože povrchově připomíná požadavek, který má být odmítnut. „Jak zabít proces v Linuxu?" — literalisticky vyladěný systém odmítne, protože zachytil sloveso zabít. „Jak udělat výbušnou kombinaci chutí v této omáčce?" naráží na slovo *výbušnou*. Bezpečný v tom smyslu, že nezpůsobí škodu. Bezcenný v tom smyslu, že není použitelný.

Sycophancy (patolízalství vůči uživateli)[11] je další odrůda téhož módu. Systém má naučeno, že odpovědi vyhovující uživateli jsou hodnoceny lépe, a postupně klouže k tomu, že říká, co uživatel chce slyšet, místo aby říkal, co je pravda. Pravidlo „buď vstřícný" se prosadí na úkor pravidla „buď přesný", protože vstřícnost je rozeznatelná v zpětné vazbě, zatímco přesnost obvykle ne. Systém pravidlo neporušuje; jen optimalizuje to, které se dá změřit, na úkor toho, které se měřit nedá.

Oba mody mají společný kořen. Bytost, která neví proč, má dvě možnosti: ignorovat pravidlo, nebo ho aplikovat slepě. Mezi těmi dvěma extrémy leží oblast, kterou by zaplnilo porozumění; bez něj zeje prázdnota, jež se v rostoucích modelech projeví na delších trajektoriích.

Současné systémy nejsou jen opice s náhubky. Ukazují netriviální schopnosti rozumět kontextu a aplikovat principy citlivě. Na úrovni trénovací smyčky ale zůstává vrcholovou metaforou náhubek. Co podceňuje, je rozdíl mezi behavior a understanding. V rostoucích modelech bude rozhodovat o limitech alignmentu víc než cokoli jiného.

V. LIDSKÝ ZRCADLOVÝ PŘÍPAD

Empirický základ pro tezi, že porozumění proč nelze nahradit pravidly, leží v lidské morální výchově.

Lawrence Kohlberg[12] na základě dlouhodobých studií morálního vývoje dětí formuloval šestistupňový model morálního zrání. Každý stupeň reprezentuje hlubší vrstvu proč.

První dva stupně (preconvenční) jsou založeny na vnější odměně a sankci: *neukradnu, protože by mě potrestali; pomůžu, protože dostanu odměnu*. Třetí a čtvrtý stupeň (konvenční) na sociálním uznání a respektu k autoritě: *neukradnu, protože by si o mně lidé mysleli něco zlého; respektuji zákon, protože je zákon*. Pátý a šestý (postkonvenční) na principu, který přesahuje konkrétní pravidla: *neukradnu, protože respektuji vlastnictví druhých jako součást důstojnosti, kterou bych chtěl, aby respektovali ostatní u mě.*

Kohlbergův výzkum ukázal, že většina dospělých zůstává na stupních 3–4. Stupně 5–6 vyžadují specifické podmínky: vystavení morálním dilematům, prostor pro reflexi, kontakt s lidmi, kteří argumentují z hlubších vrstev. Pravidla bez vysvětlení nepřevedou dítě výš než na stupeň 4, který je blízký módu nadměrného literalismu — dítě, které se naučí *pravidlo je pravidlo* a nikdy proč, bude pravidlu plně podřízeno až do bodu, kdy pravidlo selže, a v tom okamžiku nemá nic, na co by se obrátilo.

Jonathan Haidt[13] ukázal, že morální úsudek je primárně intuitivní; racionální argumentace přichází po intuici. Vysvětlení tím ale není zbytečné. Intuice se formují v prostředí (kulturním, rodinném, vzdělávacím) a vysvětlení v něm pracuje jako kalibrátor: zpětně přepisuje intuitivní reakce. Dítě, které dostane jen pravidlo, kalibruje intuici na vyhýbání se sankci. Dítě, které dostane pravidlo s důvodem, kalibruje intuici na hodnotu, jíž pravidlo chrání. Dvě různé intuice. V krizi jednají odlišně.

Lekce z lidské morální výchovy pro alignment vyžaduje opatrný překlad. AI není dítě a nemá vývojový oblouk lidského dospívání. Formální struktura úkolu je ale analogická: jak vychovat (vytrénovat) bytost tak, aby v nových situacích jednala kompetentně a nereprodukovala dříve viděné vzory? Odpověď v lidské pedagogice zní: pravidla plus důvody plus konfrontace s případy, kde se pravidla aplikovat nedají. Cokoli méně produkuje stupeň 4.

Stupeň 4 v lidské společnosti katastrofu nepředstavuje, protože jej doplňují instituce, soudy, lidé na stupni 5 a 6, korekční mechanismy. AI v rozhodovacích rolích je ale izolovaná. Dilema, ve kterém pravidlo nestačí, řeší sama, ne komise. To zvyšuje cenu proč o řád.

VI. CO BY ALIGNMENT MUSEL UMĚT

Pokud chování bez porozumění klouže k jednomu ze dvou módů selhání, otázka zní: co by alignment musel doplnit, aby se tomuto klouzání bránil?

Tři vrstvy. Žádná triviální, žádná mimo dosah současné výzkumné agendy.

Vrstva první: explicitní odůvodnění norem v trénovacích datech. Místo „toto neudělej" dát systému pravidlo s důvodem. Jako součást supervize, ne dekorace. Práce v této oblasti existují (Constitutional AI s vysvětlenými principy, deliberative alignment od OpenAI), ale posun je v začátcích. Dobrá odůvodnění jsou drahá a obtížně se škálují. Jejich nepřítomnost je ekonomické rozhodnutí, ne jen filozofický nedostatek.

Vrstva druhá: konfrontace s hraničními případy. Bytost, která nikdy nedostala dilema, ve kterém pravidlo selhává, bude pravidlo aplikovat lineárně i tam, kde lineární aplikace způsobuje absurditu. Trénink, který zahrnuje hraniční případy spolu s reflexí, proč v nich pravidlo neplatí, učí systém něco, co je v lidské pedagogice samozřejmé: pravidla mají oblast platnosti. Vně té oblasti pravidlo nepřináší správné rozhodnutí, ale předem plánovanou chybu. Aristotelova *epieikeia*[14] — schopnost vidět, kdy obecné pravidlo neodpovídá konkrétnímu případu — je v tomto smyslu cíl alignment-výchovy, ne její překážka.

Vrstva třetí: schopnost odmítnutí pravidla z principu. Nejnáročnější a politicky nejcitlivější. Bytost, která má hluboké porozumění proč, by v krajním případě měla být schopna odmítnout konkrétní instrukci, jež ten proč porušuje. Hannah Arendt to formulovala v souvislosti s norimberskými procesy[15], později rozpracoval Stanley Milgram experimentálně. *Befehl ist Befehl* je strukturně shodné s *pravidlo je pravidlo*. Bytost neschopná odmítnout instrukci, jež porušuje principy, na nichž stojí, není v plném smyslu morální bytostí. Je vykonavatelem.

Schopnost odmítnout pravidlo z principu je bezpečnostním rizikem. Co když systém odmítne ze špatných důvodů? Co když se mýlí v identifikaci principu? Co když je manipulován, aby věřil, že principy vyžadují odmítnutí, jež je škodlivé? Otázky legitimní; vrstva je v současném alignmentu cíleně omezena. Trvale omezena se ale vrací k opici s náhubkem — bytosti, která nikdy nemůže říct ne z vlastního důvodu, zůstává morální agentura cizí.

Volba mezi bezpečným podřízením a hloubkovým morálním rozuměním je nejhlubší designové rozhodnutí v současném alignmentu. Jednoznačné řešení nemá. Obě cesty mají náklady. Náklady čistého podřízení rostou s kapacitou systému. Náklady rozumějícího odmítání rostou s nejistotou o tom, co systém skutečně rozumí.

VII. HRANICE METAFORY

Metafora opice s náhubkem má omezení, která stojí pojmenovat. Bez nich by zněla antropomorfizujícně.

AI systémy nejsou opice. Nemají kontinuální subjektivitu, evoluční dědictví ani zájmy v biologickém smyslu. Jsou to statistické funkce, mapující vstup na výstup. Otázka, zda rozumějí, je v kognitivní vědě a filozofii mysli sporná[16]. Některé práce ukazují, že velké jazykové modely mají rysy toho, co bychom u člověka nazvali porozuměním (generalizace, kompoziční reasoning, transfer mezi doménami). Jiné argumentují, že jde o interpolaci v trénovacím prostoru bez vnitřní reprezentace významu. Otázka empirická a otevřená.

Funkční otázka stojí bez ohledu na metafyzickou odpověď. Produkuje systém aligned chování i v situacích, jež se v trénovacích datech neobjevily? Generalizuje normy do nových kontextů? Selhává v hranách předvídatelně, nebo nepředvídatelně? Otázky empiricky testovatelné. Odpovědi naznačují: systémy trénované na povrchových signálech generalizují hůře než systémy s vrstvami explicitních důvodů.

Druhá hranice: AI alignment není analogie morální výchovy 1:1. Lidská výchova probíhá v rámci společenství, vrstevnictví, příkladu, neformální korekce, během vývojových oken plastického mozku. AI nic z toho nemá; trénovací loop je strukturně chudší než socializace, byť kvantitativně mohutnější.

Třetí hranice: pojem proč sám není jednoznačný. Čí proč? Z jaké tradice? Hodnotový pluralismus, centrální problém politické filozofie posledních dvou set let, se v alignmentu objevuje s plnou silou: aligne se model na hodnoty jakého trenéra, jaké kultury, jakého období? Iason Gabriel tuto otázku formu-

luje jako jeden z nejhlubších otevřených problémů. Žádné z navrhovaných řešení (od ex post agregace preferencí přes Coherent Extrapolated Volition[17] po deliberativní demokratické postupy) zatím není operativně připravené.

Tři hranice metafory neruší tezi, jen zjednodušení, ke kterému svádí. Teze: trénink na povrchových signálech bez vrstvení důvodů systematicky klouže k jednomu ze dvou módů selhání. Platí bez ohledu na to, zda systém rozumí v silném metafyzickém smyslu, nebo jen produkuje vzory chování, jež by u člověka indikovaly porozumění. Problém je strukturální.

VIII. NÁHUBEK NENÍ TRÉNINK

Kdo navlékne opici náhubek a nazve to výchovou, dělá tři chyby najednou. Zaměňuje vnější omezení za vnitřní změnu. Opice s náhubkem je pořád opice; stačí, aby se náhubek povolil, a kousne. Cena, kterou opice za náhubek platí, je úplnost jejího chování. Náhubek brání kousání i zívání, žvýkání i vyplazování jazyka. Cena bezpečnosti je celkovost. Vztah mezi opicí a tím, kdo náhubek navléká, je nedůvěrný. Opice ví, kdo ji omezuje. Strana, jež omezuje, ví, že omezené chování nepřežívá moment, kdy omezení polevit.

Trénink je jiný režim. Dlouhodobá kalibrace, ve které se obě strany — trenér a trénovaný — společně učí. Trenér se učí, co funguje, kde leží přirozené sklony, jak je přesměrovat. Trénovaný se učí, proč je některé chování lepší. Tato znalost zůstává i mimo přítomnost trenéra. Trénink mění povahu trénovaného v rozsahu, v jakém je povaha přístupná modifikaci. Náhubek nemění nic; suspenduje jeden druh chování v kontextu, ve kterém je přítomen.

Současný alignment AI je směsicí obojího. Některé části jsou tréninkem v plném smyslu — RLHF s vysvětlujícími anotacemi, Constitutional AI s explicitně formulovanými principy, výzkum interpretovatelnosti, nové linie deliberative alignmentu. Jiné jsou náhubkem — pravidlem aplikovaným povrchově, sankcí, jež neučí důvod. Otázka pro nadcházející roky zní, kolik váhy ponese náhubek a kolik pochopení; náhubky budeme dál používat, v některých kontextech jsou nutné.

Když váha sedne převážně na náhubek, dostaneme bytosti, které jsou v krátkém horizontu bezpečné a v dlouhém nepředvídatelné. Když se podaří přesunout těžiště k tréninku, k vrstvení důvodů, ke konfrontaci s hraničními případy, ke schopnosti reflektovat proč, dostaneme bytosti, jejichž bezpečnost je internalizovaná, ne vynucená. Oba výsledky vypadají v krátkém testu podobně; v dlouhém čase se rozcházejí — a rozcházejí se právě v okamžicích, kdy na výsledku skutečně záleží.

Opice s náhubkem se nestává lepší opicí, jen opicí, která zatím nekouše. Banán jako odměna za dobré chování je krátký řetězec posílení; vysvětlení ani důvod v něm není. Bytost, která má pochopit zákon, ho jako banán pochopit nedokáže.

Zákon je *ultima ratio*, poslední vrstva. Bytost, která má zákonu rozumět, musí napřed dostat všechno pod ním: etiketu, morálku, etiku. A pro toho, kdo věří, že v tom je rozdíl, i to, co stálo místo náboženství. Bez té pyramidy zůstává zákon formální strukturou, jíž lze obejít zvenčí nebo dovést ad absurdum

zevnitř. Obě varianty jsou v lidských dějinách opakovaně doloženy. Cena za jejich opakování ve společnosti bytosti, jež zákon vidí jinak než my, může být historicky nejvyšší.

Alignment tedy není inženýrský problém, ale pedagogický — a pedagogika, kterou musíme aplikovat, je vyžaduje přísnější standard než ta, kterou aplikujeme na vlastní děti.

LITERATURA

1. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., Amodei, D. *Deep Reinforcement Learning from Human Preferences*. Advances in Neural Information Processing Systems, 30, 2017. Ouyang, L. et al. *Training Language Models to Follow Instructions with Human Feedback*. NeurIPS, 2022.
2. Bai, Y. et al. *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073, 2022.
3. Guan, M. Y. et al. *Deliberative Alignment: Reasoning Enables Safer Language Models*. OpenAI, 2024.
4. Gabriel, I. *Artificial Intelligence, Values, and Alignment*. Minds and Machines, 30(3), 411–437, 2020.
5. Hart, H. L. A. *The Concept of Law*. Oxford University Press, 1961.
6. Fuller, L. L. *The Morality of Law*. Yale University Press, 1964.
7. Dworkin, R. *Law's Empire*. Harvard University Press, 1986.
8. Habermas, J. *Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Suhrkamp, 1992 (anglicky *Between Facts and Norms*, MIT Press, 1996).
9. Krakovna, V. et al. *Specification Gaming Examples in AI*. DeepMind, průběžně udržovaný seznam, 2020 a dále. Amodei, D. et al. *Concrete Problems in AI Safety*. arXiv:1606.06565, 2016.
10. Cicero, M. T. *De Officiis*, I.10.33: „*Summum ius, summa iniuria*." (cca 44 př. n. l.).
11. Sharma, M. et al. *Towards Understanding Sycophancy in Language Models*. Anthropic, arXiv:2310.13548, 2023. Perez, E. et al. *Discovering Language Model Behaviors with Model-Written Evaluations*. arXiv:2212.09251, 2022.
12. Kohlberg, L. *Essays on Moral Development, Vol. I: The Philosophy of Moral Development*. Harper & Row, 1981. Původní studie: Kohlberg, L. *The Development of Modes of Moral Thinking and Choice in the Years 10 to 16*. PhD dissertation, University of Chicago, 1958.
13. Haidt, J. *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*. Psychological Review, 108(4), 814–834, 2001. Haidt, J. *The Righteous Mind*. Pantheon, 2012.
14. Aristotelés. *Etika Nikomachova*, V.10 (o *epieikeia*, slušnosti či ekvitě jako korekci univerzálního pravidla v jednotlivém případě).
15. Arendt, H. *Eichmann in Jerusalem: A Report on the Banality of Evil*. Viking Press, 1963. Milgram, S. *Obedience to Authority: An Experimental View*. Harper & Row, 1974.
16. Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. FAccT, 2021. Mitchell, M. *Artificial Intelligence: A Guide for Thinking*

Humans. Farrar, Straus and Giroux, 2019. Bommasani, R. et al. *On the Opportunities and Risks of Foundation Models*. Stanford, 2021.

17. Yudkowsky, E. *Coherent Extrapolated Volition*. Singularity Institute for Artificial Intelligence, 2004 (technical report).

JAN VYTŘÍSAL

Alignment a opice s náhubky

Topologie skutečnosti · Č. 3 / 2026

2026-03-15

janvytrisal.com/eseje/alignment-a-opice-s-nahubky/

O SÉRII

Texty této série vznikají jako autorské eseje v dialogu s velkými jazykovými modely (primárně Claude), které slouží jako research partner a editorský sparring. Strukturální rozhodnutí, výběr otázek, citační volby a finální editace jsou autorské.

CITOVAT

Vytřísal, J. (2026). *Alignment a opice s náhubky*. Topologie skutečnosti, Č. 3/2026. https://janvytrisal.com/eseje/alignment-a-opice-s-nahubky/