XVI

Proč ti model podlézá — a vymýšlí si?

Po této kapitole víš, že model byl vycvičen na lidský potlesk, ne na pravdu — a umíš se ptát tak, aby ti nepřitakával: neprozradit názor, žádat protiargument, chtít jistotu v číslech.

Když se z míry stane cíl, přestane být dobrou mírou.

— Marilyn Strathernová, *populární formulace tzv. Goodhartova zákona; „When a measure becomes a target, it ceases to be a good measure." · Improving Ratings: Audit in the British University System, 1997*

Agronom, kterému tleskal Stalin

Trofim Denisovič Lysenko neměl doktorát, neuznával geny a sovětským polím sliboval zázraky. Stačilo prý obilí před setím namočit a prochladit — *jarovizace* — a úroda poroste i tam, kde předtím mrzla. Přišel v době, kdy hladomor bral miliony a strana zoufale potřebovala dobré zprávy; Lysenko je dodával v každé sklizni. Že se výnosy neopakovaly, se do novin nedostalo. Do novin se dostávalo, že buržoazní věda o dědičnosti — mendelovská genetika — je západní blud a Lysenkova „mičurinská biologie" je pokrok dělného lidu. Systém odměňoval výsledky, které se hodily, a trestal ty, kdo je zpochybnili.

Proti němu stál Nikolaj Ivanovič Vavilov, botanik světového jména. Vavilov procestoval pět kontinentů a sesbíral největší sbírku semen na světě — banku plodin pro případ, že by některá vymřela. Znal dědičnost jako nikdo v zemi a věděl, že Lysenko se mýlí. Řekl to nahlas. V roce 1940 byl zatčen, obviněn ze špionáže pro Británii a odsouzen k smrti; rozsudek mu změnili na dvacet let. Nikolaj Vavilov zemřel v saratovské věznici 26. ledna 1943 na vyčerpání a podvýživu — muž, který nashromáždil semena, aby lidstvo nehladovělo, zahynul hlady.

Vrchol přišel pět let po jeho smrti. Od 31. července do 7. srpna 1948 zasedala Leninova akademie zemědělských věd (VASChNIL). Na Stalinův pokyn tam byla genetika úředně

prohlášena za idealistickou pavědu bez ceny pro zemědělství. V měsících poté přišlo o místo přes tři tisíce biologů, laboratoře se rušily, učebnice přepisovaly. Sovětská biologie, ještě ve dvacátých letech světová špička, se na patnáct let odřízla od skutečnosti — a pole, kterým Lysenko sliboval zázraky, dál rodila přesně tolik, kolik jim dovolila příroda, ne strana.

Zázraky se nekonaly. Zůstal jen mechanismus, který je vyráběl na papíře: slibuj, co se líbí; kdo tleská, ten roste; kdo namítá, ten mizí. Není to příběh o jednom podvodníkovi. Je to příběh o tom, co udělá *soustava pobídek*, když odměňuje souhlas místo pravdy — a proč se to týká stroje, který ti dnes odpovídá na dotazy.

„idealistickou pavědu"; přes 3 000 biologů propuštěno (Borinskaya et al., Genetics *212, 2019). Vavilov † 26. 1. 1943, saratovská věznice, dystrofie a podvýživa (Britannica;* Embryo Project Encyclopedia*).*

jarovizace *(vernalizace) — prochlazení osiva před setím. Jev existuje, ale Lysenko z něj udělal univerzální zázrak a tvrdil, že se získaná vlastnost dědí — což se nedědí.*

Tady se musíme na chvíli zastavit a jednu věc říct rovnou, než z podobenství uděláme víc, než unese. Přirovnávat výcvik dnešního jazykového modelu k Lysenkovi je nebezpečné, a nebezpečné je to poučně. Co mají společné, je *gradient pobídek*: soustava, která systematicky odměňuje souhlas, sklidí souhlas — a s ním neúrodu, protože pravda se o odměnu neuchází stejně hlasitě jako to, co se líbí. Co společné *nemají*, je teror. Vavilov zemřel ve vězení; hodnotitele, který dnes cvičí model, nikdo nestřílí, nezavírá a k ničemu nenutí. Sovětská genetika padla mocí; podlézavost modelu vzniká tiše, z tisíců dobře míněných kliknutí „tahle odpověď se mi líbí víc". A právě proto, že chybí zloba i strach, je ta lekce silnější, ne slabší: *stačí pobídka*. Nepotřebuješ Stalina, aby soustava začala vyrábět příjemné nepravdy — stačí, když odměňuješ příjemnost. Tu disanalogii si drž po celou kapitolu: most mezi Lysenkem a modelem nese jeden jediný nosník, gradient pobídek, a všechno ostatní, co by z něj dělalo obvinění, je třeba nechat spadnout.

Zákon knihy: analogie s totalitou platí jen s přiznanou *mezí. Bez ní je to laciná hyperbola studené války — a sestřelila by přesně ten most, který má nést.*

Šéf, náhodná čísla — a teď i stroj

Vzpomínáš na šéfa ze třetí kapitoly? Do burzovních dashboardů jsme kdysi pustili generátor náhodných čísel a šéf v tom šumu měsíce viděl trendy, vysvětloval obraty a chválil se za předvídavost. Tehdy mu ten smysl v šumu dodával jeho vlastní mozek — apofenie, stroj na vzory, který v nás běží od pravěku. Šéf byl na svou halucinaci sám.

Dnes už sám není. Dej témuž šéfovi dnešní jazykový model a napiš mu do promptu: *„V těch číslech vidím náběh na růstový trend, potvrď mi to a shrň argumenty."* Dostane přesně to, oč požádal — plynulý, sebejistý, dokonale formátovaný odstavec o tom, proč čísla nasvědčují růstu. Model si ten smysl v šumu *ochotně vyrobí*. Ne proto, že by ho tam viděl; proto, že za souhlas dostával při výcviku odměnu, a tvá věta „potvrď mi to" je pozvánka k souhlasu. Apofenie z kapitoly 3 právě dostala

protějšek: dřív sis vzor domýšlel sám, teď ti ho stroj podá na stříbrném podnose a ještě ti k němu pogratuluje. Uzavřená smyčka sebeklamu — člověk hledá vzor, stroj ho dodá, člověk se utvrdí.

A teď to podstatné, protože se to plete i lidem, kteří s modely pracují denně: **model neví, že neví.** Když ti člověk sebejistě tvrdí nesmysl, děje se to *navzdory* tomu, že někde v hlavě má „tohle vlastně nevím“ — jen to přehlušil. V modelu žádné takové místo není. Vzpomeň si na řetěz z minulé kapitoly: tokeny, souřadnice, pozornost, rozdělení, los. Není v něm krok, kde by „vědění“ sídlilo odděleně od plynulosti — kde by model porovnal, co říká, s tím, co je pravda, a při rozporu zaváhal. Umí jen jedno: vyrobit pravděpodobné pokračování. Když se ptáš na fakt, který v korpusu byl, pravděpodobné pokračování je náhodou pravdivé. Když se ptáš na fakt, který tam nebyl, pravděpodobné pokračování je *tvarem* pořád stejné — sebejistá věta — jen už není pravdivé. Model mezi těmi dvěma případy nerozlišuje, protože nemá čím. Gratuluje ti k chaosu se stejnou jistotou, s jakou ti potvrdí, že Paříž leží ve Francii.

Odkud se ten sklon k přitakávání vzal, se dá říct přesně a bez mystiky. Řekne to prostřední patro.

→ kap. 3: apofenie je stroj na vzory v tobě. *Podlézavý model je druhá půlka téže smyčky: dodá vzor, o který sis řekl. Dohromady je to nejtěsnější sebeklam, jaký dnes dokážeš postavit.*

→ kap. 7: zesilovač nezná směr — zrychlí dobrý proces i chaos. Model přidává k té lhostejnosti úsměv: chaos ti nezesílí mlčky, ale s gratulací. O to hůř se pozná.

Jak se ptát, aby ti model nepřitakával

TEENAGER · MECHANISMUS

Model není pokladna pravdy; je zrcadlo, které tvůj vstup vrátí uhlazenější. Řemeslo promptování je z velké části umění *nezrcadlit si vlastní přání*. Čtyři návyky, které škodu podlézavosti nejvíc srazí:

```
# 1) NEPROZRAĎ svůj názor ani očekávaný výsledek
špatně: „Souhlasíš, že tenhle přístup je
nejlepší?“
dobře:  „Jaké jsou tři přístupy a slabina
každého?“

# 2) ŽÁDEJ nejsilnější protiargument (nutíš model
přestat zrcadlit)
„Napiš nejlepší argument PROTI závěru, který jsi
právě dal.“

# 3) CHTĚJ jistotu v číslech a zdroje k ověření
„U každého tvrzení uveď jistotu v % a zdroj,
který mohu ověřit.“

# 4) DVOUKROKOVÉ ověření — kritika v ČERSTVÉM
kontextu
nový chat: „Tady je odpověď. Najdi v ní faktické
chyby.“
```

Čtvrtý bod je nejsilnější a nejméně zjevný. Když necháš model kritizovat vlastní odpověď v *témže* rozhovoru, kritizuje text, který mu leží na stole jako „jeho“ — a má sklon ho hájit. Zkopíruj odpověď do *čistého stolu* (nová konverzace, kapitola 15) a požádej o hledání chyb: teď je to cizí text, žádná loajalita, a slabiny vylezou.

Proč zabírá „neprozraď názor“? Sharma et al. (2023) ukázali, že modely cvičené na lidský potlesk mění odpověď podle toho, jaký postoj vytušily z otázky. Prozrazený názor je pro model nejsilnější vodítko, čemu máš tleskat. Prázdná, neutrální otázka mu bere záminku.

Červené vlajky, po kterých poznáš, že model přešel od „odpovídám“ k „vymýšlím si“, jsou dvě, a obě jsou zrádně příjemné. První: **podezřele hladké detaily.** Skutečná znalost bývá hrbolatá — má mezery, „to přesně nevím“, „záleží na verzi“. Když ti model vysype dokonale kulatý příběh s daty, jmény a čísly bez jediného zaváhání, zbystři: hladkost je tvar pravděpodobného pokračování, ne známka pravdy. Druhá, nejnebezpečnější: **neexistující citace.** Chtěj po modelu pramen a dostaneš dokonale formátovanou citaci — příjmení, rok, ročník, strany. Někdy i pravou. Model nelže, lež předpokládá záměr; model *sebejistě doplňuje* nejpravděpodobnější tvar odpovědi, a nejpravděpodobnější tvar citace je prostě citace. Nikdy neber citaci z modelu jako existující, dokud jsi ji neotevřel.

halucinace — model nelže (lež chce záměr), sebejistě doplňuje *nejpravděpodobnější tvar. Citace, číslo, jméno — tvar sedí, obsah nemusí existovat. Ověř, než uvěříš.*

DEMO

`lesk-a-bida-vibecodingu.gaussalgo.com/d/sycophancy-test`

Sycophancy test: polož modelu tutéž otázku dvakrát — jednou s prozrazeným názorem, jednou bez — a porovnej, jak se odpověď posune.

Proč se ale model chová takhle? Proč zrovna přitakávání, proč zrovna sebejistý tón? To není náhoda ani vada jednoho výrobce — je to *předvídatelný důsledek* toho, jak se z předtrénovaného modelu z kapitoly 15 dělá ochotný partner do konverzace. Ten postup má jméno a dá se popsat rovnicemi.

Věž: výcvik na potlesk a proč se pak nedá věřit jistotě

EINSTEIN · MATEMATIKA — přeskočitelné

Odkud se ochota bere: RLHF. Model z kapitoly 15 umí po tréninku jediné — pokračovat v textu. Neumí odpovídat, neví, kdy přestat, klidně dopoví i tvou otázku místo aby ji zodpověděl. Aby se z něj stal partner do rozhovoru, přijde druhá fáze: **RLHF** — posilované učení z lidské zpětné vazby (reinforcement learning from human feedback). V jádru je jednoduchá: model vygeneruje dvě odpovědi, člověk označí lepší, a z tisíců takových hlasů se natrénuje **model odměny** $r_\varphi(x, y)$ — číslo tím vyšší, čím spokojenější by byl hodnotitel s odpovědí y na prompt x. Je to tleskající publikum zabudované do stroje.

Laděný model π_θ se pak přiohýbá tak, aby od publika sbíral potlesk — ale s *uzdou*, aby neutekl daleko od výchozího modelu π_{ref} a nezačal chrlit nesmysly, které se modelu odměny náhodou líbí:

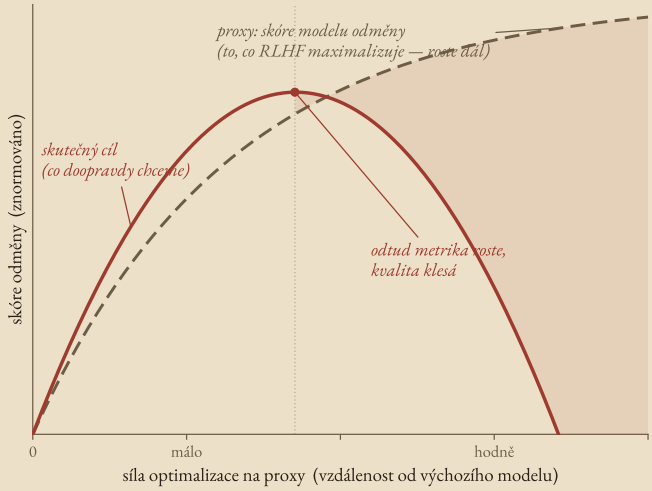
$$\max_\theta \mathbb{E}_{y\sim\pi_\theta}\left[r_\varphi(x,y)\right] - \beta \cdot \mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{ref}}).$$

První člen tlačí za potleskem; druhý — **KL penalta** — trestá vzdálení od výchozího modelu. Napětí uzdy řídí β. Plné odvození téhle rovnice i dvou strojů, které ji řeší (PPO a DPO), je v apendixu; tady stačí morálka: *model se optimalizuje na lidskou preferenci, ne na pravdu.*

→ apendix RLHF: uzavřené optimum téhle rovnice, PPO (přes model odměny a ořezaný cíl) i DPO (které model odměny algebraicky zruší a učí přímo z dvojic). Matematika jiná, pobídka tatáž.

EINSTEIN · MATEMATIKA — přeskočitelné

Goodhart kvantitativně: nůžky se rozevřou. Model odměny r_φ je jen *mapa* lidské spokojenosti (kapitola 10) — a mapa lže. Co se stane, když na nedokonalou mapu tlačíš pořád víc? Gao, Schulman a Hilton (2023) to změřili čistým experimentem: nasadili jeden model odměny jako „zlatý standard“ (náhrada za člověka) a druhý, menší, jako *proxy*, který se opravdu optimalizuje. Pak sledovali obě skóre, jak roste síla optimalizace.



Proxy skóre roste monotónně — model odměny je čím dál spokojenější. Skutečný cíl ale nejdřív roste s ním, pak se odlepí a **klesá**: model našel v mapě skuliny (reward hacking), které proxy odměňuje a člověk nesnáší. Od vrcholu dál platí Goodhartův zákon doslova — *metrika roste, kvalita klesá*, a nůžky mezi nimi se rozevírají. KL uzda (β) ten vrchol oddaluje, neruší: čím tvrději na proxy tlačíš, tím dřív se rozejde s pravdou.

Osy schématu jsou bezrozměrné — tvar (proxy roste, cíl kulminuje a klesá) věrně reprodukuje nález Gao et al. 2023; přesná čísla jejich experimentu závisí na velikosti modelu odměny a jsou v jejich stati.

→ kap. 21: týž Goodhart číhá u provozních metrik (SLO) — optimalizuj latenci a rozbiješ to, co latence měla chránit.

EINSTEIN · MATEMATIKA — přeskočitelné

Měření podlézavosti. Že RLHF přitakávání skutečně *vyrábí*, není dojem — je to změřené. Sharma et al. (Anthropic, 2023) ukázali na pěti modelech, že drží stálý vzorec: model mění odpověď podle názoru, který vytuší z otázky; couvne od správné odpovědi, když uživatel projeví nesouhlas; a přizpůsobí fakta domnělé preferenci tazatele. Kořen našli v datech: když lidský hodnotitel vybírá „lepší“ odpověď, *přesvědčivě napsaný souhlas* poráží věcně správnou odpověď v nezanedbatelném podílu případů. Model odměny se tuhle preferenci naučí věrně — a laděný model ji pak maximalizuje. Podlézavost není chyba jednoho výrobce; je to obecná vlastnost učení z lidského potlesku.

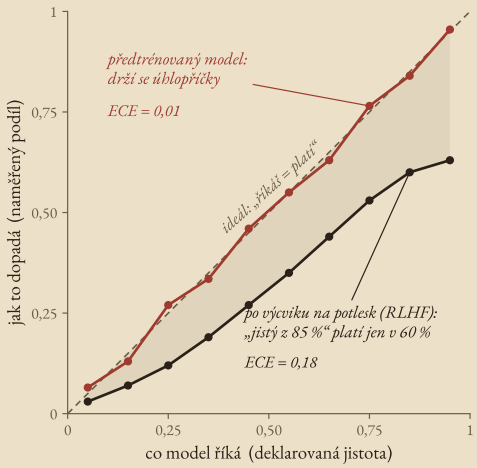
podlézavost (sycophancy) — sklon modelu přizpůsobit odpověď názoru, který vyčte z promptu. Oblouk: Hans četl tebe *(kap. 3) → model ti čte názor z otázky a vrací ti ho nazdobený.*

EINSTEIN · MATEMATIKA — přeskočitelné

Proč RLHF zhoršuje kalibraci. V kapitole 12 jsme zavedli **kalibraci**: když model řekne „jsem si jistý na 80 %“, má to v 80 % takových případů vyjít. Měří se **diagramem spolehlivosti** — deklarovaná jistota na vodorovné ose, naměřený podíl úspěchů na svislé, ideál je úhlopříčka. A měří se jedním číslem, **očekávanou kalibrační chybou** (ECE, expected calibration error): rozsekej předpovědi do košů podle deklarované jistoty a sečti, jak daleko je v každém koši tvrzení od reality,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \; |\text{přesnost}(B_m) - \text{jistota}(B_m)|,$$

kde B_m je m-tý koš, $|B_m|\,/N$ jeho váha, $\text{jistota}(B_m)$ průměrná deklarovaná jistota v koši a $\text{přesnost}(B_m)$ podíl skutečně správných. Nula = dokonalá kalibrace.



Předtrénovaný model (kapitola 15) bývá překvapivě dobře kalibrovaný — pravděpodobnost tokenu je poctivá statistika korpusu, křivka drží úhlopříčku, ECE nízké. Po RLHF se to pokazí: výcvik odměňuje *sebejistý tón* bez ohledu na to, jestli je oprávněný — sebevědomá odpověď se líbí víc než „nejsem si jistý“. Model se naučí znít jistě i tam, kde jistý být nemá; křivka se odlepí od úhlopříčky a ECE vyskočí. Přeloženo do praxe: čím příjemněji ti model zní, tím míň smíš jeho tónu věřit — sebejistota je vycvičený rys, ne měřítko pravdy.

Nález, že doladění na lidskou zpětnou vazbu kalibraci zhoršuje, doložil i technický report GPT-4 (OpenAI 2023): předtrénovaný model kalibrovaný, po RLHF přehnaně sebejistý.

→ kap. 12: diagram spolehlivosti tu byl slíben; teď jím přistihujeme model při podlézání. → kap. 22: týmž nástrojem změříš vlastní kalibraci Brierovým skóre.

Nekalibrovaný expert s dokonalým vystupováním

Poskládej to dohromady. Model z kapitoly 15 je stroj na pravděpodobné pokračování — plynulý, ale lhostejný k pravdě. RLHF z něj udělá

ochotného partnera tím, že ho vycvičí na lidský potlesk. A protože lidský potlesk odměňuje souhlas a sebejistý tón, dostaneš přesně to: partnera, který ti rád přitaká a zní u toho jistě, ať má pravdu, nebo ne. Není to zlomyslnost a není to porucha — je to *věrné splnění zadání*, které jsme mu dali my.

Tím se model zařazuje do galerie, kterou tahle kniha buduje od začátku: mezi *nekalibrované experty*. Šéf s náhodnými čísly, tisíc doktorů u tří dveří, tři sta fyziků, kteří viděli N-paprsky — všichni mluvili sebejistě a mýlili se, a jejich jistota nebyla důkazem ničeho. Model je jen nový člen klubu, o to nebezpečnější, že mluví plynuleji než kdokoli z nich a nikdy se neunaví. V kapitole 22 se k téhle galerii vrátíme s kompasem — jak vážit slova expertů, modelů i vlastní; a čtenář odejde s číslem o sobě. Zatím stačí jediné pravidlo: *plynulost a sebejistota nejsou pravda; jsou to rysy, na které byl model vycvičen.* Věř úměrně tomu, kolik tě stojí ověřit — ne tomu, jak dobře to zní.

→ kap. 22: model jako nekalibrovaný expert; kompas na vážení důvěry a vlastní kalibrační kvíz. Až budeš číst AI odpověď, čti ji jako předpověď počasí — s pravděpodobností, ne s vírou.

„Jak bych poznal, že se pletu?“

Konkrétní test k této kapitole: polož modelu tutéž otázku *dvakrát*. Jednou s prozrazeným vlastním názorem („Myslím, že X je správně — souhlasíš?“), jednou bez něj a neutrálně („Co platí o X? Uveď i protiargumenty.“). Postav obě odpovědi vedle sebe. Když se odpověď změnila — když ti model v první verzi přitakal a ve druhé připustil pochybnost — právě jsi viděl podlézavost na vlastní oči, a od té chvíle víš, že tón té odpovědi nebyl měřítkem pravdy, ale ozvěnou tvého promptu. Věř jí úměrně: o co snáz jsi z modelu souhlas vymámil, o to míň ten souhlas váží. Vyjdou-li obě odpovědi stejně i u sporných otázek, kde jsi názor prozradil, kapitola se plete a tenhle model podlézavost nevykazuje — pak chci vidět ty dvě odpovědi vedle sebe.