

VIII

Proč se přivázat
ke stěžni?

Kratší zastavení mezi dvěma dějstvími. Po něm víš, proč „budu opatrný, prohrává — opatrnost je rozhodnutí v reálném čase, a přesně tam tě píseň dostane. A umíš sepsat svůj první pakt: pravidla, která si dáš dřív, než otevřeš chat. Zbytek knihy je návod, jak ten stěžen postavit.

Chceš-li je slyšet ty sám, ať k stěžni tě přivážou pevně, rukama, nohama zkraje, a lana ať k stěžni ti uvážou. A budeš-li prosit a poroučet, aby tě rozvázali, tím pevněji tehdy ať přitáhnou ještě víc provazů na tě.

— Kirké Odysseovi, *Homér, Odysseia, 12. zpěv (rada, jak minout Sirény).*
Přebásněno podle překladu O. Vaňorného.

Smlouva s budoucím já

Odysseus se plavil domů a věděl, že ho čeká úžina, kde zpívají Sirény — a že jejich píseň není lest ani past nastražená na hlupáky. Je *krásná*. Tak krásná, že každý, kdo ji uslyší, se k ní vrhne, zapomene na domov, ženu i posádku a zůstane na tom ostrově, dokud nezetlí v hromadě kostí, které tam po předchozích okouzlených leží. Nešlo o to, že by námořníci před ním byli slabší nebo hloupější. Ta píseň prostě přemluví každého. To je celý vtip Sirén: neselháváš proto, že jsi špatný, ale proto, že ony jsou neodolatelné.

A Odysseus udělal něco, co z něj v téhle knize dělá prvního inženýra: přiznal si, že *v tu chvíli* neobstojí. Že až uslyší tu píseň, bude prosit, křičet, poroučet — a každý argument, který v tom okamžiku vymyslí, aby ho pustili, bude znít rozumně. Vůle v reálném čase byla ztracená předem. Tak rozhodl dřív, než k té úžině dorazil.

Ten plán, dodejme poctivě, nevymyslel sám — poradila mu ho kouzelnice Kirké, u níž předtím rok pobýval, a Odysseus měl jen tolik rozumu, že poslechl. Rada zněla ve dvou krocích. Posádce zalep uši voskem, aby píseň vůbec neslyšeli a mohli veslovat dál. A sebe — protože ty ji slyšet *chceš* — nech přivázat ke stěžni a přikaz

mužům, ať tě za žádnou cenu neodvazují, ať budeš říkat cokoli; a budeš-li prosit o rozvázání, ať tě přitáhnou lany ještě pevněji.

Tak propluli. Posádka s ucpanýma ušima klidně veslovala. Odysseus slyšel tu nejkrásnější píseň světa, poznal ji celou — a když se v okouzlení začal rvát z pout a řval na muže, ať ho pustí, oni ho podle rozkazu jen přivázali víc. Píseň dozněla, loď byla dávno z dosahu a Odysseus byl jediný člověk, který Sirény slyšel a přežil. Ne proto, že by měl silnější vůli než ti v hromadě kostí. Proto, že rozhodl *předem* — a odebral svému budoucímu, okouzlenému já možnost to rozhodnutí zvrátit.



Ten příběh není o starověké plavbě. Je o jediné technice, na které stojí celá druhá polovina téhle knihy, a proto stojí za to ho vzít vážně. Filozof Jon Elster z té scény roku 1979 vytáhl přesný pojem: *sebeomezení* — rozhodnutí, kterým dnešní já svazuje ruce zítřejšímu já, protože ví, že zítřejší já bude slabší. Neříká „snaž se víc,“. Říká něco chytřejšího: „*odeber si možnost selhat*“. A není to jen filozofie. Medicína z Odyssea udělala právní nástroj — pacient, který zná svou nemoc, může podepsat závaznou listinu, že až přijde krize a on bude odmítat pomoc, personál ho má léčit i proti jeho tehdejší vůli. Říká se jim, samozřejmě, Ulyssovy smlouvy.

Oko: píseň je skutečně krásná

Musíme začít u toho, co většina varování před technologií zamlčuje, protože bez toho zbytek nedává smysl. Píseň Sirén je krásná — a jazykový model je skutečně užitečný. To není metafora, kterou tady mírním, abych ti stroj nezošklivil. Je to doslova pravda. To, co dnes model napíše za vteřinu, mě kdysi stálo noci; poradí ti, vysvětlí, vygeneruje, doplní — a dělá to dobře, rychle a lichotivě. Kdyby ta píseň nebyla krásná, nikdo by tuhle kapitolu nepotřeboval. Slabé pokušení se překoná mávnutím ruky. Silné pokušení, to, které je *opravdu dobré*, se nepřekoná vůlí v tu chvíli — a přesně o takovém je řeč.

A tady je ta věta, kvůli které je celá coda v knize: *proto nefunguje „budu opatrný“*. Opatrnost je rozhodnutí učiněné v reálném čase — ve chvíli, kdy píseň zpívá, kdy máš odpověď na dosah, kdy stačí zmáčknot enter a mít hotovo. A v reálném čase prohráváš, protože přesně na ten okamžik je pokušení navrženo. „Zkontroluju to potom.“ „Tenhle jeden krok přeskočím, vypadá to správně.“ „Nasadím to a testy dopíšu, až bude čas.“ Každá z těch vět zní ve chvíli, kdy ji říkáš, naprosto rozumně — stejně rozumně jako Odysseův řev, ať ho pustí. Slabé místo není

stěžeň (a vosk do uší) — vědecká metoda jako pakt s budoucím já. Nosná metafora celé knihy; křtí se tady a vrací se v refrénu každé kapitoly až po závěr. Kdo píseň slyšet nepotřebuje, smí si zacpat uši — tenhle čtenář ji slyšet chce, proto stěžeň.

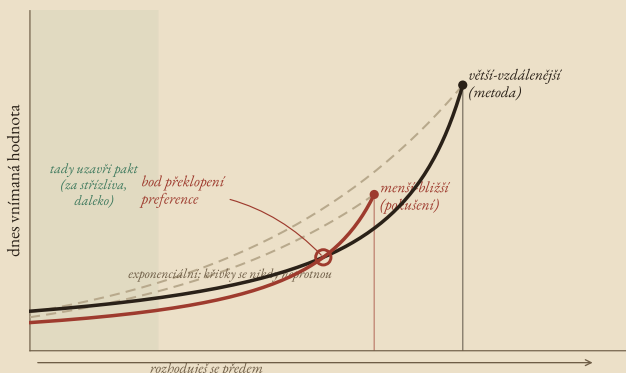
EINSTEIN

Jon Elster, *Ulysses and the Sirens* (Cambridge UP, 1979): systematická teorie sebeomezení (precommitment). Precommitment zkoumal už dřív ekonom Thomas Schelling; Elster jej vztáhl na slabou vůli a nedokonalou racionalitu. „Ulyssovy smlouvy“, v psychiatrii jsou reálné: v Nizozemsku právně závazné od r. 2008 (zákon Wvvgg od 1. 1. 2020).

tvá inteligence. Slabé místo je spoléhat se v reálném čase na to, že se rozhodneš správně, když je krása na dosah.

Funguje jen jedno: *závazek uzavřený předem*. S posádkou, s procesem, se sebou samým — dřív, než píseň začne. To je celý rozdíl mezi Odysseem a hromadou kostí na ostrově. Ne odolnost v okamžiku, kdy už je pozdě, ale rozhodnutí učiněné za střízliva, daleko od úžiny, kdy tvá preference ještě míří ke správné věci. A pak ruce svázané tak, aby je okouzlené já nemohlo rozvázat.

Podívej se ještě na tu posádku, protože v ní je schovaná lekce, kterou memová verze příběhu přehlíží. Námořníci si zacpali uši voskem — píseň vůbec neslyšeli — a klidně proveslovali. A to bylo *v pořádku*. Nebyli to zbabělci ani hlupáci; měli práci (dostat loď domů) a k té píseň nepotřebovali. Kdo píseň slyšet nepotřebuje, smí si legitimně zacpat uši. Ve světě AI to znamená: kdo svou práci zvládne bez modelu a nechce ho, má plné právo si ho nepouštět k tělu. Vosk je čestná strategie. Není to prohra.



Hyperbolické diskontování (Atndie): z dálky ceníš víc větší-vzdálenější odměnu — metodu. Jak se blíží pokušení, hodnota menší-bližší odměny prudce vysočší a v bodě překlopení přeroste metodu — preference se obrátí. Exponenciální diskontování (dárkování) se nikdy neprotne. Proto se rozbíhajících předem: vlevo, kde metoda ještě vede — ne upravo, kde už ji prohráváš.

Ozvěna kap. 7: „buď opatrný“, je jako doufat, že se lidé v reply-all bouři budou chovat rozumně. Lavinu neporazíš přáním, porazíš ji číslem nastaveným předem. Pakt je totéž pro svou vlastní vůli: nastavené pravidlo místo naděje na sebekázeň v tu chvíli.

Proč vůle v reálném čase prohraje: z dálky ceníš víc větší-vzdálenější odměnu (pocitovou metodu). Jak se pokušení blíží, hodnota menší-bližší odměny (hotovo teď) prudce vyskočí a v bodě překlopení přeroste metodu — preference se obrátí. Proto pakt uzavíráš vlevo, daleko od úžiny.

Ale tahle kniha není pro toho, kdo si ucpal uši. Je pro toho, kdo píseň slyšet *chce* — kdo chce ten model používat, protože je vážně dobrý, a nechce si o tu krásu dát zacpat uši. A pro toho vosk není odpověď. Pro toho je odpověď stěžň: slyšet celou píseň, užít si ji naplno — a přitom mít ruce svázané pravidly, která nedovolí okouzlenému já poslat loď na skály. Odysseus je hrdina téhle knihy právě proto, že si vosk nezacpal. Chtěl slyšet. A tak si postavil něco lepšího.

Tři reakce na jednu píseň

Shrňme, co tím máme, protože jsou to přesně tři možné postoje ke stroji a víc jich není. Před sirénou stojíš a máš tři volby, ne dvě.

Tři reakce na sirény. Píseň je skutečně krásná — a ke každé z reakcí patří někdo, pro koho je správná.

1. VOSK DO UŠÍ — píseň neslyšíš vůbec
= odmítnutí AI. Legitimní! Kdo model k práci nepotřebuje,
má právo si ho nepouštět. Posádka
doveslovala domů.
riziko: míjíš nástroj, který ti mohl znásobit práci.
2. SKOK DO MOŘE — vrhneš se za písni bez pout
= slepý vibe coding. "Však ono to nějak dopadne."
Bereš, co model dá, nasazuješ neověřené,
věříš dojmu.
riziko: hromada kostí na ostrově. Tohle je ta past.
3. STĚŽEŇ — slyšíš celou píseň, ale ruce máš
svázané předem
= metoda navržená za střízliva. Užiješ si
nástroj naplno
A neztroskotáš, protože pravidla rozhodla
dřív než pokušení.
tahle kniha staví STĚŽEŇ.

Skoro každý průšvih s AI, který uvidíš i způsobíš, je skok do moře v přestojení — ne zlá vůle, ne hloupost, jen rozhodnutí v reálném čase, kdy píseň zpívala a ruce byly volné. „Vypadalo to správně, tak jsem to poslal dál.“ „Fungovalo to na mém příkladu, tak jsem to nasadil.“ To nejsou selhání charakteru. Jsou to selhání *protokolu* — chyběl stěžně. A stěžně se, na rozdíl od charakteru, dá postavit. Prkno po prkně, a začínáme hned tady.

Kolo: tvůj první pakt

Přejdeme od příběhu k rukám, protože „přivaž se ke stěžni“, je krásný obraz a k ničemu, dokud nevíš, co konkrétně jsou ta lana. Lano je *pakt*: pravidlo, které si dáš dřív, než otevřeš chat, a které tě zavazuje bez ohledu na to, co ti píseň zazpívá. Není to dobré předsevzetí („budu pečlivější“). Je to tvrdý, předem daný uzel — a čím dřív ho uvážeš, tím pevněji drží.

Pakty vibe codera. Každý z nich je rozhodnutí přesunuté z okamžiku pokoušení do klidu předem. Uzavři je dřív, než otevřeš chat.

□ TEST PŘED GENEROVÁNÍM

Napiš, jak poznáš správný výstup, DŘÍV než ho model vyrobí.

Potom už jen zaškrtněš – nesvádíš se s hotovou odpovědí.

□ COMMIT PŘED KAŽDÝM EXPERIMENTEM

Ulož pozici dřív, než začneš zkoušet. Návrat je zdarma,

tak si můžeš dovolit odvahu. (→ kap. 18: commit = mikropakt.)

□ BUDGET PŘEDEM – čas i peníze

"Dám tomu 90 minut / 5 dolarů na tokeny, pak končím."

Číslo dané předem, ne "ještě chvilku" v jednu ráno.

□ DEFINICE HOTOVO

Napiš, co znamená "hotovo", než začneš. Jinak je hotovo

tam, kde tě přestane bavit se dívat – a to je moc brzy.

□ VĚTA "CO MĚ PŘESVĚDČÍ, ŽE JE TO ŠPATNĚ"

Napsaná DŘÍV, než se zeptáš modelu. Protože potom už

budeš hledat důvody, proč je odpověď dobrá.

Všimni si, co mají všechny ty pakty společného, protože v tom je celý trik a je to tentýž trik jako Odyseův. Ani jeden se nespolehá na tvou vůli *ve chvíli*, kdy model zazpívá. Každý přesouvá rozhodnutí do klidu *předem* — kdy ještě nemáš hotovou odpověď před očima, kdy tě netlačí čas, kdy tvá preference míří ke správné věci. Test napsaný dřív než kód nejde ohnout, aby vyhovoval tomu, co model náhodou vyrobil. Věta „co mě přesvědčí, že je to špatně“, napsaná dřív než dotaz, tě chrání před tím, aby ses okouzleně chytal důkazů, že je to dobře. To není nedůvěra ke stroji. Je to nedůvěra k vlastnímu okouzlenému já — a ta je namístě, protože ono je slabé přesně tam, kde je píseň nejkrásnější.

A tady je most, který drží celou knihu pohromadě, protože ta pravidla ti nespadla z nebe — poznáš je z předchozích stránek. „Věta, co mě přesvědčí, že je to špatně“, je refrén, který jsi u konce každé kapitoly už četl a od deváté ho pokřtíme naplno: *jak bych poznal, že se pletu?*

☹ EINSTEIN

Proč zabírá „předem“. George Ainslie ukázal, že hodnotu budoucí odměny diskontujeme hyperbolicky, ne exponenciálně: bližší odměna má pak neúměrně vysokou váhu. Křivky obou odměn se proto protnou — z dálky vede poctivá metoda, těsně před pokoušením ho hodnota „hotovo teď“ přeskóčí a preference se překlopí. Pakt je tab, kterým z rozhodovacího stromu budoucího já vyškrtněš tu horší větev, dokud na ni ještě nemá chuť.

„Commit před experimentem“ je disciplína levného zahazení z kapitoly o dřině — a v kapitole osmnácté dostane jméno a stroj času. „Test před generováním“, je smlouva s budoucím já, kterou v kapitole devatenácté rozebereme do posledního šroubu. Coda tyhle nitě sbírá do jedné dlaně a dává jim společné jméno: pakt.

Most: co teď postavíme

Skončeme, kde další dějství začíná, protože tahle kapitola je zároveň dveřmi. Řekli jsme, že jediná obrana proti krásné písni je závazek uzavřený předem — stěžeň, ke kterému se přivážeš. Ale zatím je to jen slovo a jeden příběh. Zbytek téhle knihy je návod, jak ten stěžeň doopravdy postavit, prkno po prkně, protože stěžeň se nedá koupit ani vydupat vůlí — musí se *vyrobiť z metody*.

A metoda má součástky, které teď jednu po druhé nasadíme. Model — abys věděl, s čím máš vlastně tu čest, když ti stroj zpívá (kapitola desátá). Měření — abys poznal, že se pleteš, dřív než tě to bude stát draho (kapitola devátá to pokřtí, dvanáctá přidá hloupého soupeře). Data, která lžou, i když čísla sedí (třináctá). Stroj, který ti podlézá, protože ho k tomu vycvičili (šestnáctá). Každá z těch kapitol je jedno prkno stěžeň. Až je přibijeme, budeš mít k čemu se přivázat — a budeš tu písni moct poslouchat naplno, protože ruce budeš mít svázané pravidly, ne strachem.

■ DEMO

lesk-a-bida-vibecodingu.gaussalgo.com/d/sireny

Tři strategie u jedné písni naživo: zvol vosk, skok do moře, nebo stěžeň a sleduj výplatu i riziko každé z nich napříč mnoha plavbami. Vedle: generátor paktu — vyplň pravidla svého příštího projektu, vytiskni si je a podepiš dřív, než otevřeš chat. První uzel uvážeš rovnou tady.

→ kap. 9–16: dějství VĚDEC staví stěžeň — metoda, měření, hloupý soupeř, chyby jako palivo.

→ kap. 14: tam si splníš úplně první pakt — zapíšeš předpověď dřív, než spustíš simulaci.

→ kap. 22: kompas na chvíli, kdy lana povolí a stěžeň nestučí.

„Jak bych poznal, že se pletu?“

Coda nemá test, kterým bys ji shodil — má úkol, protože je to začátek, ne závěr. Napiš si *ted'*, za strážlivá, dřív než otevřeš příští chat, jednu jedinou větu a doplň ji konkrétně: „*Tenble projekt zabodím — nebo nenasadím — pokud ...*“. Ne mlhavě („pokud nebude dobrý“), ale ověřitelně: pokud neprojde tímhle testem, pokud to bude stát víc než tolik, pokud nedokážu vysvětlit, proč to funguje. To je tvůj první uzel na laně — rozhodnutí přesunuté z okamžiku pokušení do klidu předem. A od téhle věty dál už jen přibíjíme prkna: stěžeň, ke kterému tu větu uvážeš, stavíme celý zbytek knihy.