Setkání umělé inteligence s přirozenou hloupostí

V roce 2012 byla řada zdravotně postižených lidí v americkém státě Idaho informována o tom, že jejich finanční podpora v rámci systému Medicaid byla omezena.[22] Ačkoli měli všichni na dávky patřičný nárok, stát snížil jejich výši – bez varování – až o 30 procent[23] a nechal je zoufale shánět peníze na svou léčbu. Nejednalo se o politické rozhodnutí; byl to výsledek nového „rozpočtového nástroje", který zavedlo ministerstvo zdravotnictví a sociální péče v Idaho – softwaru, který automaticky vypočítal úroveň podpory, kterou by měla každá osoba dostat.[24]

Problém spočíval v tom, že se zdálo, že rozhodnutí rozpočtového nástroje nedává žádný smysl. Zvnějšku by se klidně dalo říci, že vygenerovaná čísla byla v podstatě náhodná. Někteří lidé dostali více peněz než v předchozích letech, zatímco jiní zjistili, že se jejich rozpočet snížil o desítky tisíc dolarů, což je vystavilo riziku, že budou muset opustit své domovy a žít v ústavní péči.[25]

Protože nedokázali pochopit, proč byly jejich dávky sníženy, ani snížení efektivně zpochybnit, obrátili se obyvatelé na Americký svaz pro občanské svobody (American Civil Liberties Union; ACLU) s prosbou o pomoc. Jejich případ převzal Richard Eppink, právní ředitel divize v Idaho,[26] který na svém blogu v roce 2017 uvedl: „Myslel jsem si, že to bude jednoduchá záležitost zeptat se státu: Oukej, řekněte nám, proč tyto částky o tolik dolarů klesly?" Ve skutečnosti to znamenalo čtyři roky, čtyři tisíce stěžovatelů a hromadnou soudní žalobu, nežli jsme došli k jádru věci."[28]

Eppink a jeho tým začali žádostí o sdělení podrobností toho, jak algoritmus funguje, ale tým Medicaid odmítl jeho výpočty vysvětlit. Argumentovali tím, že software, který případy vyhodnotil, je chráněn coby „obchodní tajemství" a nemůže být zveřejněn.[29] Naštěstí soudce, jenž soudnímu řízení předsedal, nesouhlasil. Rozpočtový nástroj, který vládl takovou mocí nad obyvateli, byl poté odhalen a ukázalo se, že – se nejednalo

o nějakou sofistikovanou AI, o žádný elegantně vystavěný matematický model, ale o prostou tabulku v Excelu.[30]

Výpočty v tabulce byly pravděpodobně založené na skutečných případech z minulosti, ale údaje byly strašně neúplné a plné chyb, které byly z větší části zcela zbytečné.[31] Ještě horší bylo, že jakmile se týmu ACLU podařilo rozluštit rovnice, objevili „fundamentální statistické nedostatky ve způsobu, jakým byl samotný vzorec strukturován". Rozpočtový nástroj prakticky vzato účinně produkoval pro obrovský počet lidí zcela náhodné výsledky. Tento algoritmus – lze-li mu tak vůbec říkat – byl tak mizerné kvality, že soud nakonec rozhodl o jeho protiústavnosti.[32]

Můžeme zde vysledovat dvě paralelní vlákna chyb člověka. Za prvé, tuto hloupou tabulku někdo napsal; za druhé, ostatní jí naivně důvěřovali. „Algoritmus" byl ve skutečnosti jen odfláknutá lidská práce zabalená do kódu. Proč tedy lidé pracující pro stát velmi vehementně bránili něco tak hrozného?

Toto si o celé záležitosti myslí Eppink:

> Je to právě onou předpojatostí, kterou všichni máme vůči výsledkům práce počítačů – nezpochybňujeme je. Když počítač něco vygeneruje – když máte statistika, který se podívá na nějaká data a vytvoří vzorec – prostě tomu vzorci důvěřujeme, aniž bychom se ptali „hele, počkej vteřinku, jak tohle vlastně funguje?"[33]

Samozřejmě si uvědomuji, že rozbor matematických vzorců, aby se vidělo, jak fungují, nepředstavuje pro většinu lidí oblíbenou kratochvíli (i když pro mne ano). Nicméně Eppink vyzdvihl neuvěřitelně důležitý bod spočívající v naší lidské ochotě brát algoritmy jako hodnotné, aniž bychom přemýšleli o tom, co se děje za kulisami.

Během let, kdy jsem pracovala jako matematička s daty a algoritmy, jsem došla k přesvědčení, že jediným způsobem, jak objektivně posoudit, zda je algoritmus důvěryhodný, je dostat se k jádru toho, jak funguje. Podle mých zkušeností jsou

algoritmy spíše podobné kouzelnickým iluzím. Zpočátku se zdá, že nejsou ničím jiným než skutečnou magií, ale jakmile se dozvíte, jak se daný trik provádí, záhada se vypaří. Často se jedná o cosi směšně jednoduchého (nebo znepokojivě lehkomyslného), co se skrývá za plentou. Takže v následujících kapitolách se vám ze všech sil pokusím za pomoci algoritmů, které prozkoumáme, poskytnout možnost alespoň na chvilku se podívat do zákulisí. Stačí, abyste zjistili, jak se dělají triky – i když k tomu, abyste je sami prováděli, to nestačí.

Avšak dokonce i nejtvrdohlavější fanoušci matematiky se stále mohou dostat do situace, kdy algoritmy vyžadují, abyste jim slepě důvěřovali a skočili do prázdna. Možná proto, že stejně jako v případě výsledků vyhledávání systémy Skyscanner nebo Google, není možné provést dvojitou kontrolu jejich práce. Anebo je možná algoritmus, stejně tak jako onen rozpočtový nástroj v Idaho a další, s nimiž se setkáme, považován za „obchodní tajemství“. Anebo, jako u některých způsobů strojového učení, logické procesy odehrávající se v pozadí prostě nelze vysledovat.

Přijdou chvíle, kdy budeme muset postoupit kontrolu neznámému, a to ačkoli víme, že algoritmus je schopen chybovat. Chvíle, kdy budeme nuceni zvažovat náš vlastní úsudek proti úsudku počítače. Abychom se rozhodli důvěřovat svým instinktům namísto svých jeho výpočtů, to bychom museli být velmi odvážní.

Kdy se postavit na odpor

Stanislav Petrov byl ruský vojenský důstojník pověřený sledováním jaderného systému včasného varování, který chránil sovětský vzdušný prostor. Jeho úkolem bylo okamžitě upozornit nadřízené, pokud by počítač zaregistroval nějaké známky amerického útoku.[34]

Petrov byl ve službě i dne 26. září 1983, kdy krátce po půlnoci začaly houkat sirény. To byl poplach, kterého se všichni obávali. Sovětské družice odhalily nepřátelskou raketu, která mířila na ruské území. Studená válka byla v plném proudu,

takže útok se jevil vcelku věrohodně, ale přesto Petrova něco zastavilo. Nebyl si jist, zda algoritmu důvěřuje. Bylo zjištěno pouze pět střel, což se zdálo na první útok Američanů jako nelogicky malá salva.[35]

Petrov ve svém křesle strnul. Bylo to na něm: vyhlásit poplach a poslat svět do téměř jisté jaderné války; anebo počkat a ignorovat protokol, ačkoli věděl, že každá sekunda, která uplyne, znamená méně času pro vládu jeho země k zahájení protiútoku. Naštěstí pro nás všechny zvolil Petrov druhou možnost. Neměl žádný způsob, jak se ujistit, že se poplach rozezněl omylem, ale po 23 minutách – což se v takové chvíli jeví jako věčnost – když bylo zřejmé, že na ruskou půdu žádné jaderné rakety nedopadly, si nakonec uvědomil, že měl pravdu. To algoritmus chyboval.

Kdyby se systém choval zcela autonomně, aniž by jako konečná instance působil člověk jako Petrov, historie by byla nepochybně vypadala poněkud jinak. Rusko by téměř jistě zahájilo to, co by považovalo za odvetné opatření, a touto akcí spustilo plnohodnotnou jadernou válku. Pokud z tohoto příběhu plyne nějaké poučení, pak takové, že kritickou součástí procesu je právě lidský element: že mít člověka s právem veta, který prověřuje návrhy algoritmu před rozhodnutím, je jediný rozumný způsob, jak se vyhnout chybám.

Koneckonců, pouze lidé cítí tíhu odpovědnosti za svá rozhodnutí. Algoritmus pověřený komunikací s Kremlem by o možných důsledcích takového rozhodnutí naopak vůbec nepřemýšlel. Ale takový Petrov? „Velmi dobře jsem věděl, že nikdo nebude schopen napravit mou chybu, pokud nějakou udělám.“[36]

Jediný problém tohoto závěru je, že lidé nejsou vždy tak spolehliví. Někdy je správné, když se algoritmu postaví na odpor, tak jako Petrov. Ale mnohdy je nejlepší své instinkty ignorovat.

Dám vám další příklad z oblasti bezpečnosti, v níž jsou příběhy o lidech nesprávně odporujících algoritmu naštěstí vzácné – přesto je jím právě to, co se stalo během neblaze proslulé nehody na horské dráze Smiler v Alton Towers, největším zábavním parku ve Velké Británii.[37]

V červnu 2015 byli povoláni dva inženýři, aby opravili poruchu na horské dráze. Po vyřešení problému vyslali na trať prázdný vozík, aby otestovali, že všechno funguje – ale uniklo jim, že se nevrátil. Zkušební vozík se z nějakého důvodu skulil dolů ve svahu a zastavil se uprostřed dráhy.

Mezitím, bez vědomí inženýrů, zaměstnanci dráhy připravili k soupravě vozík, aby se vypořádali s prodlužujícími se frontami. Jakmile dostali z kontrolní místnosti informaci, že závada je odstraněna, začali do vozíků usazovat veselé cestující, připoutali je a první vlak poslali na trať, aniž by cokoli věděli o uvízlém prázdném vozíku vyslaném inženýry, který stál přímo v cestě.

Naštěstí konstruktéři horské dráhy s takovou situací počítali a jejich bezpečnostní algoritmy zafungovaly přesně podle plánu. Aby se zabránilo jisté srážce, byl plný vlak zastaven v horní části prvního stoupání a v řídicí místnosti se spustil alarm. Ale inženýři – přesvědčeni, že dráhu právě opravili – usoudili, že automatický varovný systém chybuje.

Překonat algoritmus nebylo snadné: oba museli souhlasit a současně stisknout tlačítko, aby horskou dráhu znovu spustili. Když tak učinili, poslali vlak plný lidí ze svahu dolů vstříc přímému nárazu do uvízlého vozíku. Následky byly strašné. Několik lidí utrpělo těžká zranění a dvě dospívající dívky přišly o nohy.

Oba tyto scénáře, otázky života či smrti, Alton Towers a Petrovův poplach, slouží jako dramatické ilustrace mnohem hlubšího dilematu. Kdo – nebo co – by měl mít konečné slovo v mocenské rovnováze mezi člověkem a algoritmem?

Boj o moc

Tento spor má dlouhou historii. Paul Meehl, profesor klinické psychologie na univerzitě v Minnesotě, rozčílil v roce 1954 celou generaci lidí, když publikoval *Klinické versus statistické předpovědi*, přičemž se jasně přiklonil k jedné ze stran sporu.[38]

Meehl ve své knize systematicky porovnával výkonnost lidí a algoritmů v celé řadě oblastí – předpovídání všeho možného

od známek studentů až po výsledky vyšetření psychiatrických pacientů – a dospěl k závěru, že matematické algoritmy, bez ohledu na to, jak jednoduché, budou téměř vždy předpovídat lépe nežli lidé.

Mnoho dalších studií v průběhu následujícího půl století Meehlovy poznatky potvrdilo. Pokud váš úkol zahrnuje nějaký druh výpočtu, vsaďte své peníze vždy na algoritmus: při určování lékařských diagnóz či prognóz prodeje, předpovídání množství pokusů o sebevraždu nebo míry spokojenosti s kariérou, a při posuzování čehokoli, od způsobilosti k vojenské službě po předpokládané akademické výkony.[39] Stroj nebude dokonalý, ale dát člověku veto nad algoritmem by znamenalo jen přidat další chybu.*

Možná by nás to nemělo překvapovat. Nejsme kalkulačky. Když jdeme do supermarketu, nenajdeme tam řadu pokladníků upřeně sledujících naše nakupování a snažících se odhadnout, kolik bude zboží stát. Naopak, máme (neuvěřitelně jednoduchý) algoritmus, který to pro nás spočítá. A povětšinou by bylo lepší přenechat to stroji. Je to, jako když se mezi piloty aerolinií říká, že nejlepší letecký tým tvoří tři složky: pilot, počítač a pes. Počítač má za úkol řídit letadlo, pilot má za úkol krmit psa. A pes má za úkol kousnout člověka, kdyby se pokusil sáhnout na počítač.

Ale náš vztah ke strojům je paradoxní. Zatímco máme tendenci příliš se spoléhat na něco, čemu nerozumíme, jakmile zjistíme, že algoritmus může dělat chyby, projeví se také poněkud nepříjemný zvyk, a to že zareagujeme přehnaně, kompletně jej odmítneme a vrátíme se zpět k našemu vlastnímu vadnému úsudku. Vědci to nazývají algoritmickou averzí. Lidé jsou méně tolerantní k chybám algoritmu, nežli ke svým vlastním – dokonce i když jsou jejich vlastní chyby větší. Tento

* Zajímavé je, že jako vzácná výjimka z převahy algoritmů co do výkonu vychází z výběru studií provedených v pozdních padesátých a šedesátých letech „diagnostika“ (jejich slovy, nikoli mými) homosexuality. V těchto příkladech lidský úsudek vedl k mnohem lepší předpovědi a překonával cokoli, co by mohl zvládnout algoritmus – což naznačuje, že existují určité věci tak intimně lidské, že data a matematické vzorce s nimi budou vždycky zápolit.