

Klasická testová teorie

PSYb2590: Základy psychometriky | Přednáška 6

25. 3. 2025 | Petra Hubatka



V minulých dílech jste viděli

1. **Substantivní teorie** charakterizující měřený atribut a proces během měření
2. **Instrumentalizace** (instrumentalization) = konstrukce nástroje, který bude k zachycení atributu citlivý
3. **Škálování** (scaling) = proces, kterým z raw dat vytvoříme informativní reprezentaci
4. **Model měření** (measurement model) = aplikace formálního (matematického) modelu, kterým otestujeme naše hypotézy o procesu měření a kvantifikujeme nejistotu spojenou s výstupy měření, např. CTT, IRT, FA, ...

V minulých dílech jste viděli

$$y_i = \mu + \lambda_1 * \theta_i + \lambda_2 * \theta_i + \dots \lambda_n * \theta_i + \varepsilon$$

$$y_{ij} = \mu_j + \lambda_j * \theta_i + \varepsilon_j$$

$$\Sigma = \Lambda \Phi \Lambda' + D_\psi$$

THE OWLS ARE NOT
WHAT THEY SEEM



Co nás dnes čeká?

Teorie chyb (*Theory of Errors*)

Klasická testová teorie (*Classical Test Theory*, CTT)

Divadlo pana Browna

Reliabilita konceptuálně

Teorie chyb (Theory of Errors)

Každé měření je zanesené nějakou chybou.

Ta může být

- systematická
- náhodná

pozorované = „pravé“ + chyba











pozorovaná vzdálenost = „pravá“ vzdálenost + chyba

konstanta

$$d_i = d_\tau + e$$

Měření vzdálenosti: Bodový odhad

1. n -krát (na našem příkladu z předchozích slidů by $n = 5$) změříme nějakou vzdálenost d .
2. Každé měření (manifestní proměnná) bude mít nějaký index, které značí jeho pořadí: $d_1, d_2, \dots, d_n$ (v našem příkladu by to bylo d_1, d_2, d_3, d_4, d_5).
3. Při odkazování se na obecně jedno měření (z řady několika měření), lze zobecnit do zápisu d_i , kde za i můžeme dosadit číslo od 1 do n a referovat tak k danému konkrétnímu měření (manifestní proměnné).
4. Vypočítáme bodový odhad (očekávanou hodnotu) vzdálenosti pomocí průměru z těchto měření: $\frac{d_1 + d_2 + \dots + d_n}{n}$.
5. Víc matematicky můžeme výše uvedené zapsat jako $E(d) = \hat{d} = \frac{\sum_{i=1}^n d_i}{n}$.

Měření vzdálenosti: Standardní chyba měření

Standardní chyba tohoto měření (SEM, Standard Error of Measurement, σ_e):

1. Pro jediné měření: $SEM = SD_d$, kde SD_d je výběrová směrodatná odchylka pozorovaných hodnot d_i .
2. Pro průměr z n měření: $SE = \frac{SD_d}{\sqrt{n}}$ -> čím víc se blížíme s počtem měření k nekonečnu, tím více je SE pro průměr z n měření bližší 0 (tím větší máme jistotu, že náš bodový odhad průměru, a tedy „pravé“ vzdálenosti je správný).

**Where am I?
And how can I leave?**

Motivace k dnešnímu tématu

Představte si, že už jste v praxi a chcete zjistit, jak je na tom Vaše klientka se schopností se soustředit. Administrujete jí proto test pozornosti. Zjistíte, že v něm dosáhla T-skóre 40.

Co můžeme o pozornosti této klientky říct v porovnání s populací? (populační průměr je 50, SD 10)

Můžeme interpretovat pouze tento skór?

Jak je na tom teda naše klientka s pozorností ve skutečnosti?

LORD and NOVICK
**STATISTICAL
THEORIES OF
MENTAL
TEST
SCORES**

WITH CONTRIBUTIONS BY
A. BIRNBAUM



Klasická testová teorie

Teorie chyb posloužila jako základ pro **Klasickou testovou teorii** (CTT)

CTT je:

- **model měření** – aplikace matematické formalizace pro modelování konkrétního procesu měření v psychologii (zde hlavně modelování chyb)
- tradičně nejrozšířenější, ačkoliv v poslední době ustupuje **Teorii latentních rysů** (již znáte!), příp. se používá ve své rozšířené formě zvané **Teorie zobecnitelnosti**.
- kvůli zaměření na chybu měření především **teorií chyby měření** (např. Sijtsma, 2020)
- **antirealistická** (není pro ni relevantní existence měřených atributů)

Teorémy CTT

1. $\tau_i = E(X_i)$ (na úrovni jednotlivce)
2. $E_i = X_i - \tau_i$ (na úrovni jednotlivce)

Hrubý skór X_i = pozorovaný výsledek měření pro člověka i

True skór (pravý skór) τ_i = průměr nekonečně mnoha měření pro člověka i

Chyba e = náhodná (tedy s ničím nekorelující), normálně rozložená chyba s $\mu = 0$

Jinými slovy:

- Pozorovaný skór je pravým skórem, který byl náhodně vychýlen chybou.
- Chyba jde do plusu i do mínusu stejně často (protože je normálně rozložená).
- Zprůměrováním nekonečna pozorování se proto efektu chyby zbavíme a zůstane pravý skór.

Klasická testová teorie

Pravý skór v CTT není definován jako hodnota měřeného atributu, ale jako očekávaná hodnota (průměr) nekonečně mnoha měření. Je to proto ryze matematická konstrukce bez vztahu k realitě.

$$X_i = \tau_i + e \text{ (klasická testová teorie)}$$

$$x_{ij} = \lambda_j * \theta_i + e \text{ (teorie latentních rysů)}$$

V podstatě totožné vzorce (pro $\lambda = 1$), ale:

1. Vzorec CTT platí pro součtový skór, zatímco vzorec TLR platí pro jednu položku.
2. Naprosto jiné filosofické pozadí (antirealismus vs. realismus).

Klasická testová teorie

CTT tedy „imituje“ opakované měření v přírodních vědách

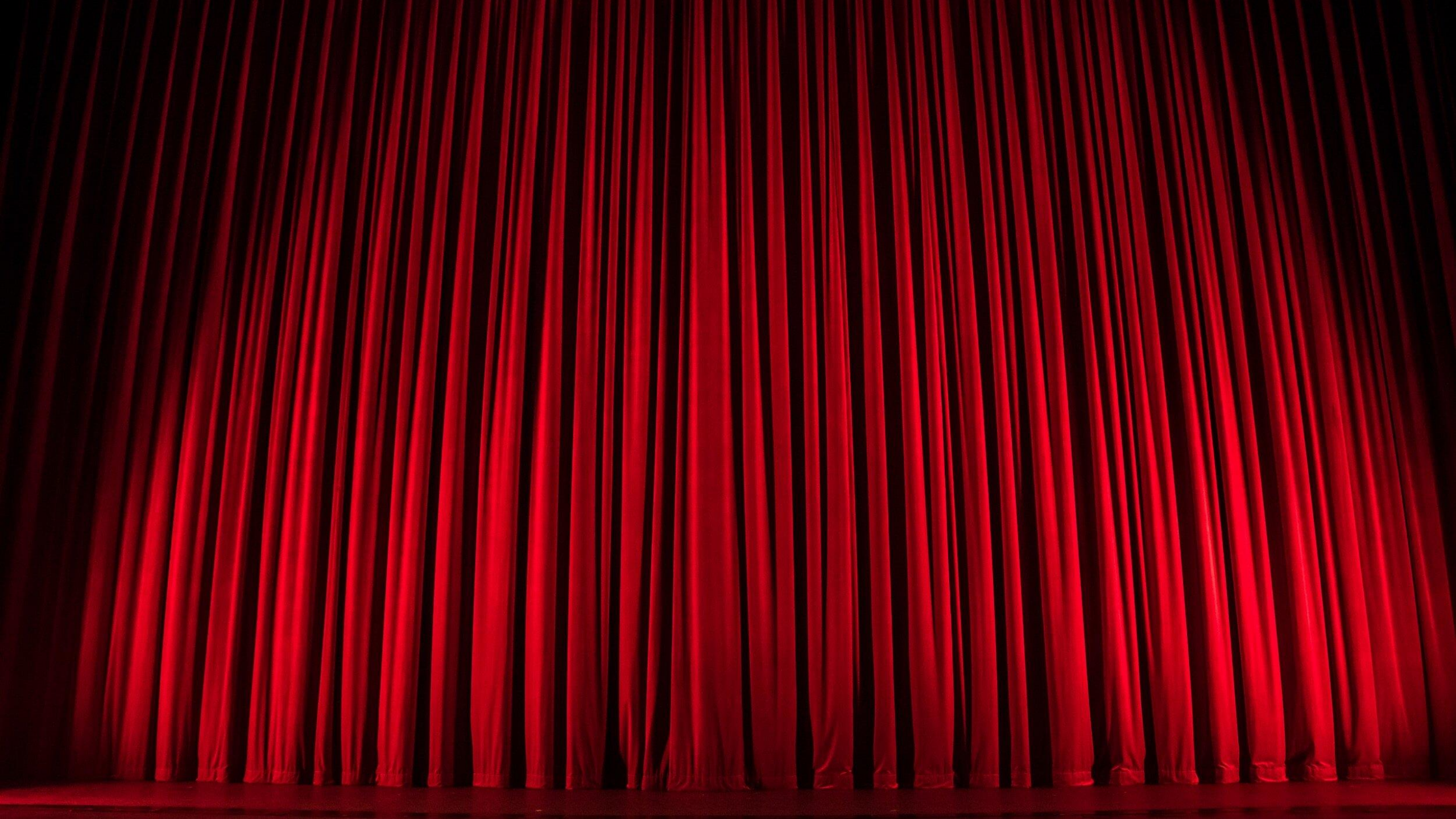
Kde ale v psychologii narážíme na problémy, chceme-li měřit opakovaně?

It is clear, therefore, that classical test theory (a) **is not concerned with series of measurements**, and (b) **could not concern itself with such series in the first place, because actual repeated measurements cannot be expected to conform to the assumptions of the theory**. Still, the syntactical formulation of the theory uses the expectation operator at an essential point in the development of the theory – namely in the definition of its central concept, the true score. **What is to be done about this awkward situation?** (Borsboom, 2005, s. 36)

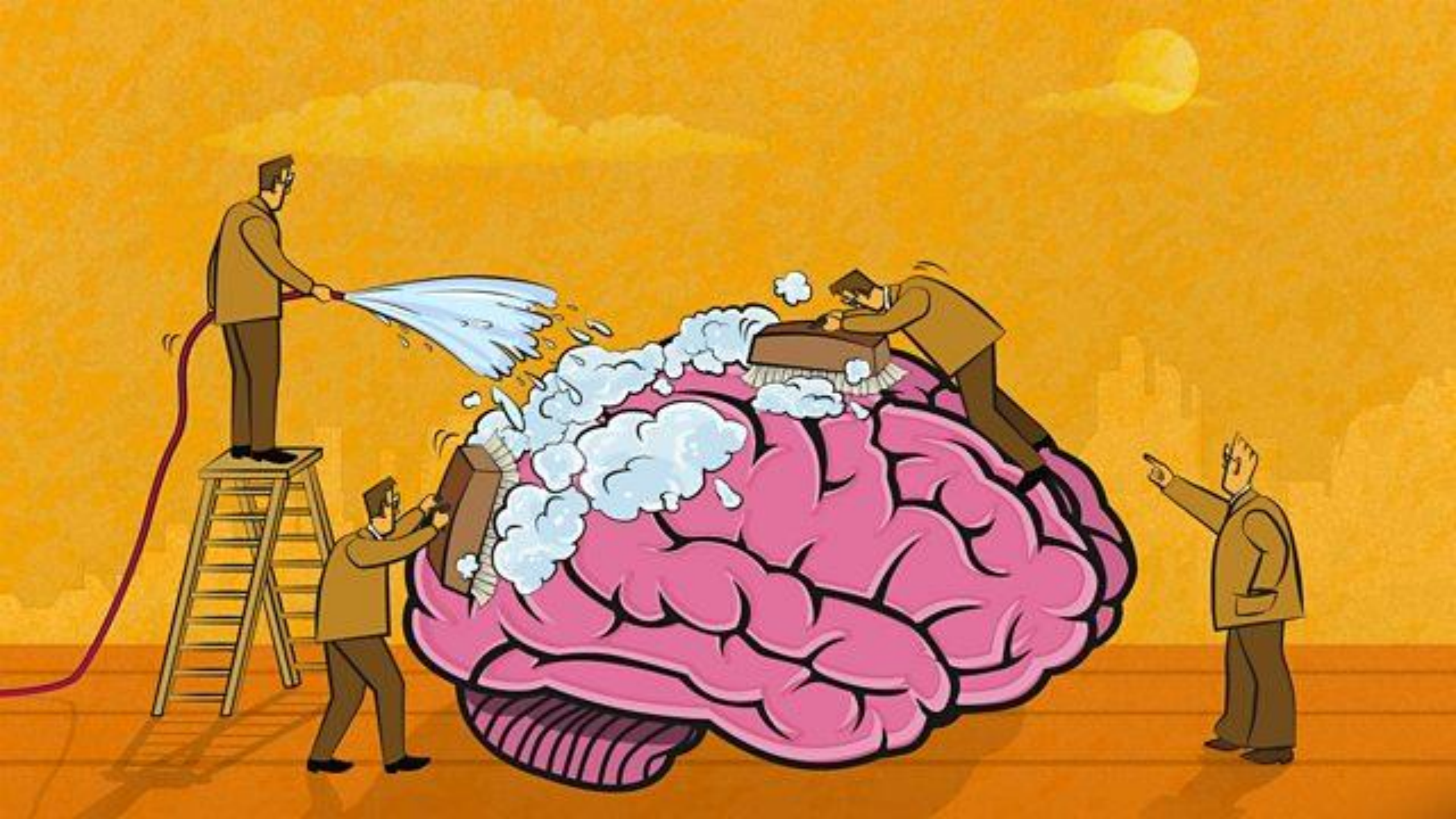
Divadlo pana Browna

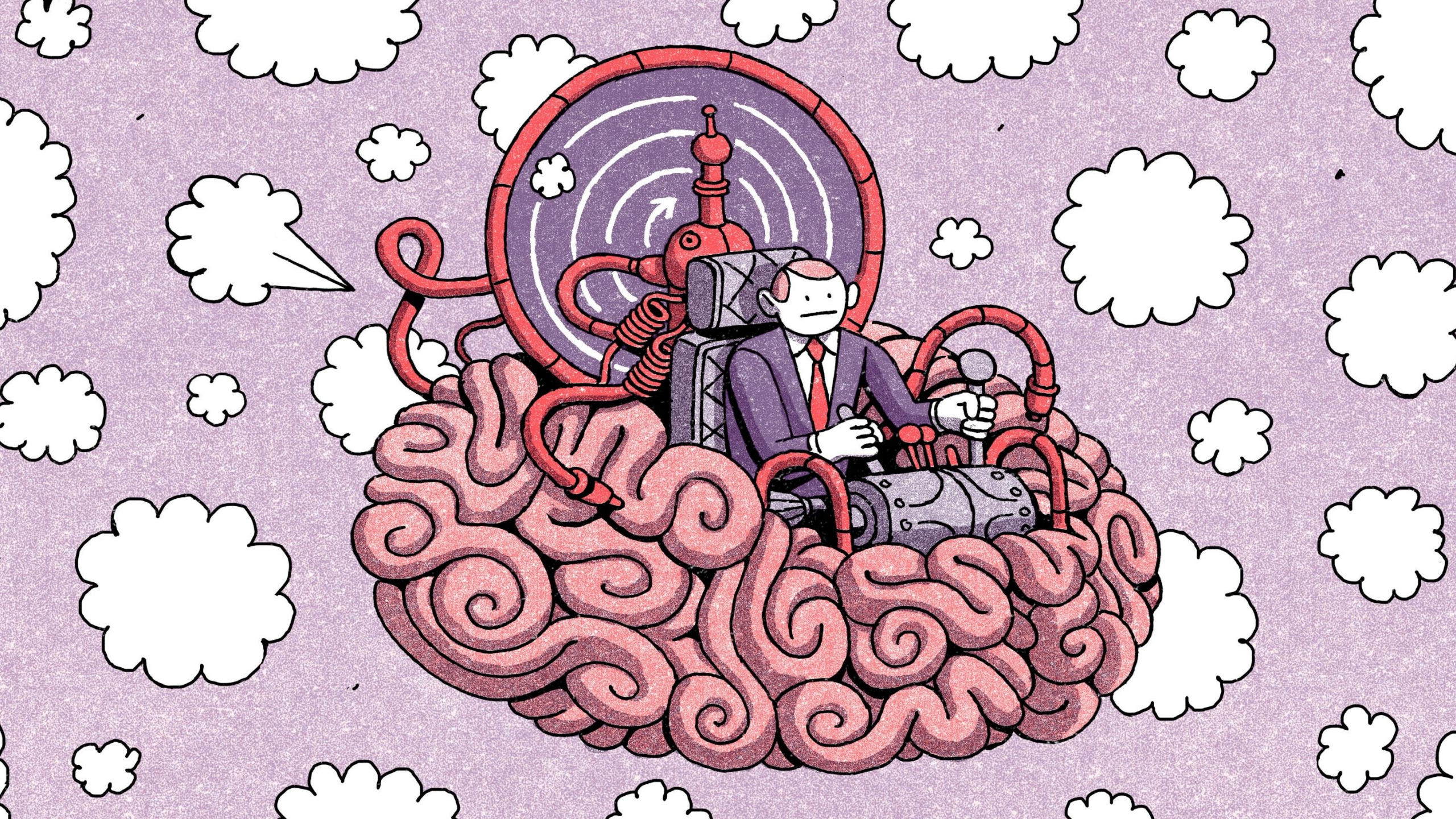
Disclaimer: Následující část může obsahovat značné prvky cringe. Doporučujeme povahám citlivým vůči cringi, aby účast na této části konzultovali s odborníkem. Vyučující se zříkávají zodpovědnosti za pocit trapnosti, které následující minuty mohou vyvolat.

Během představení prosím vypněte své mobilní telefony.

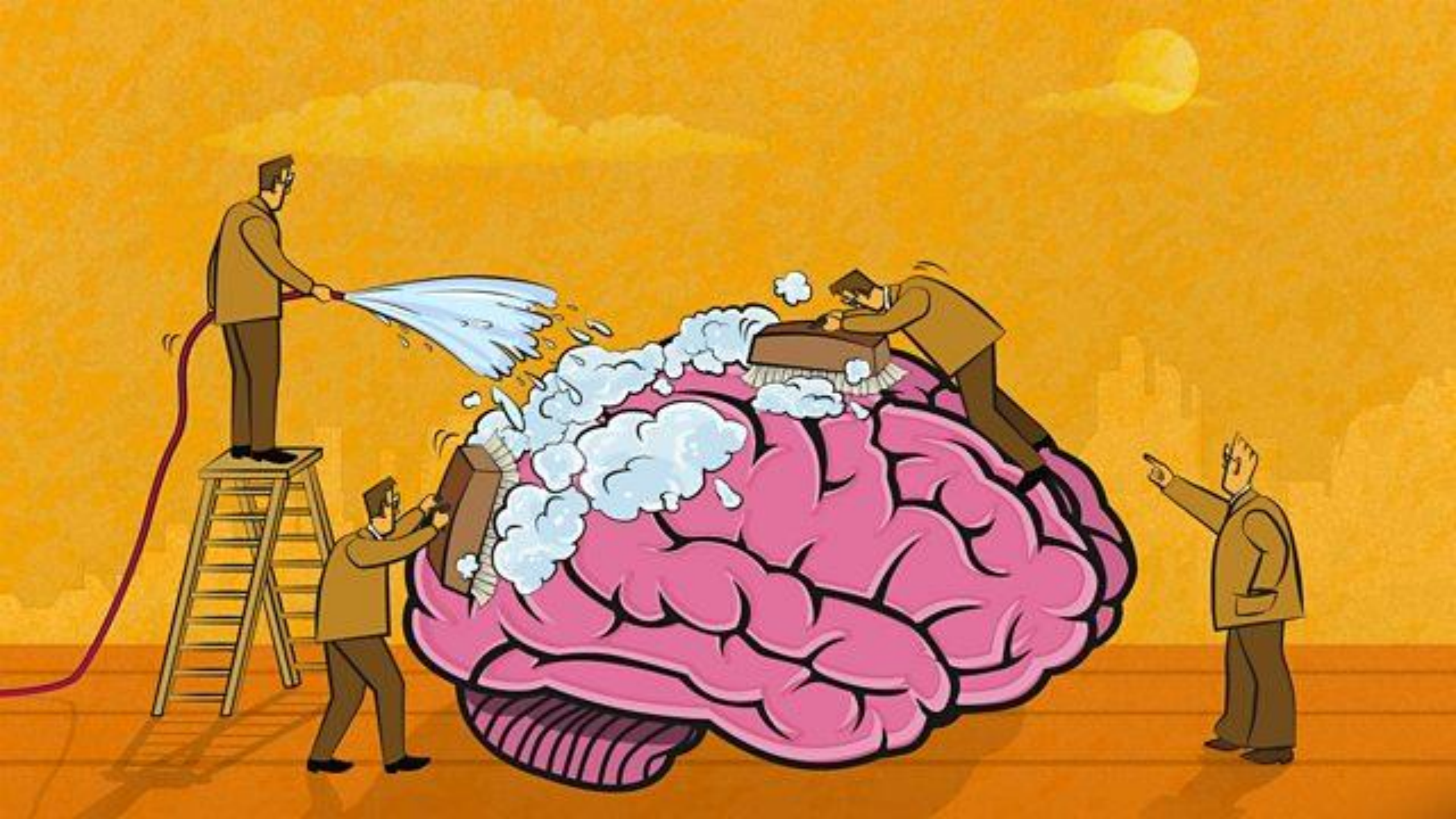


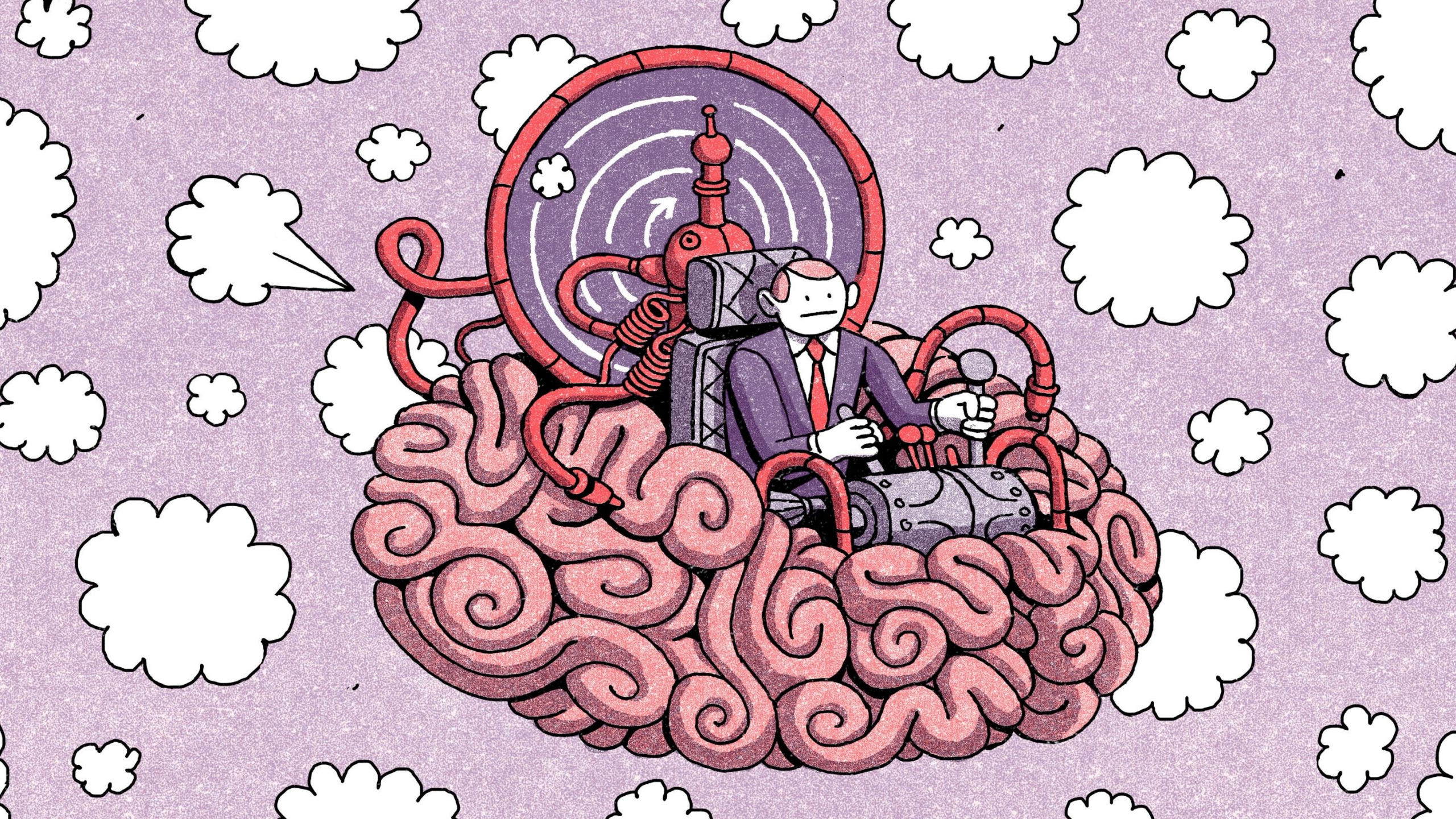




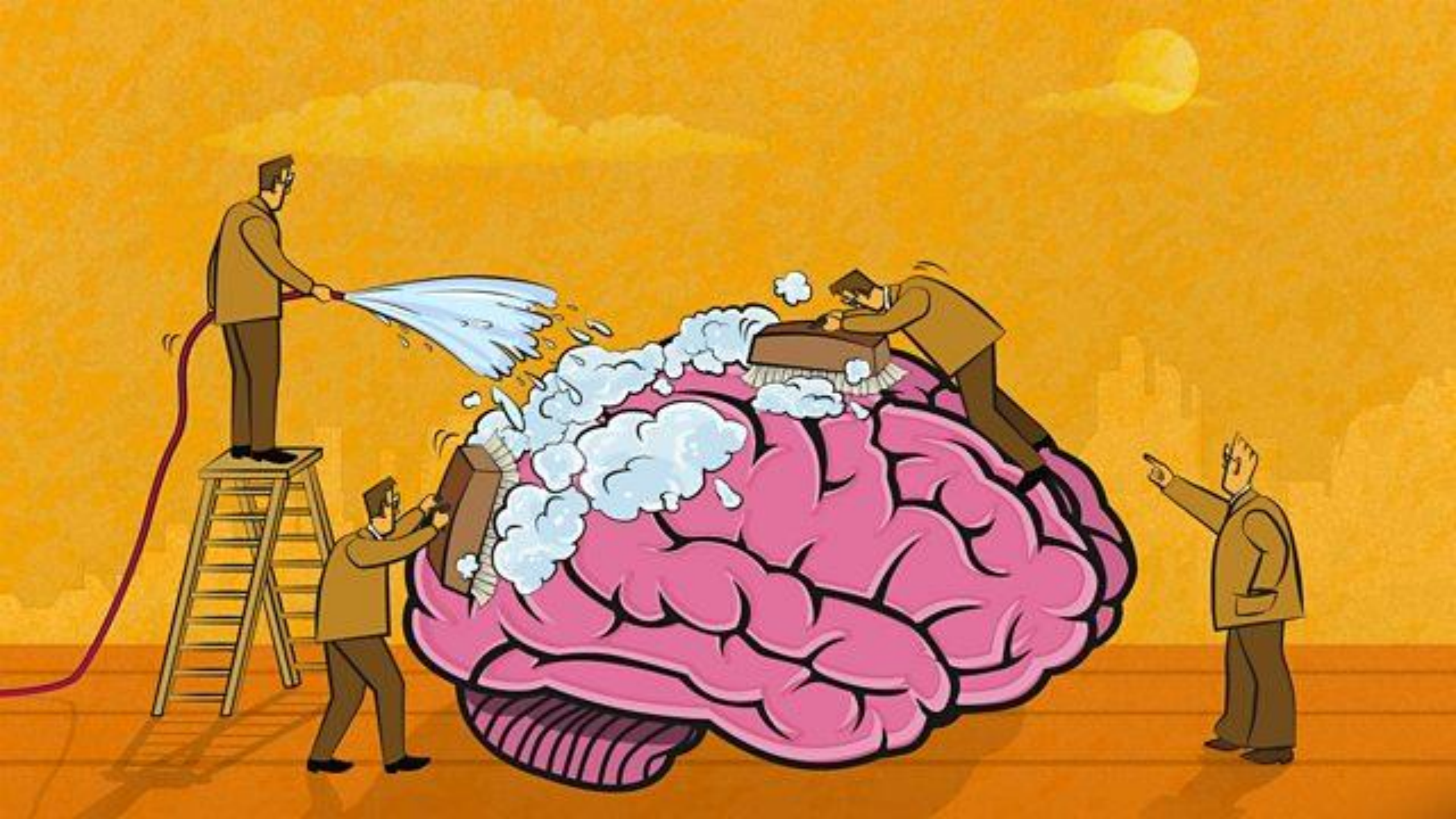


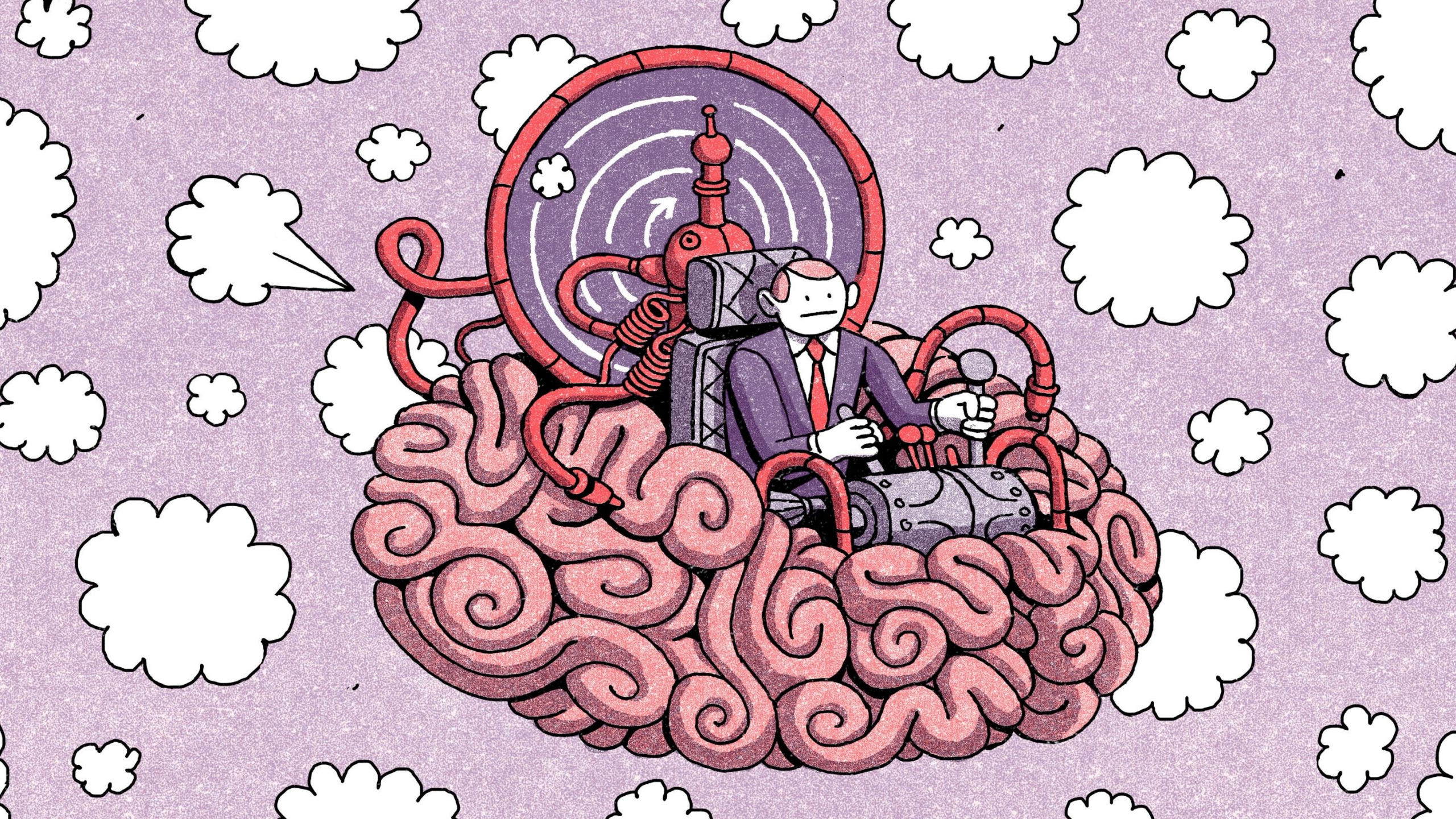




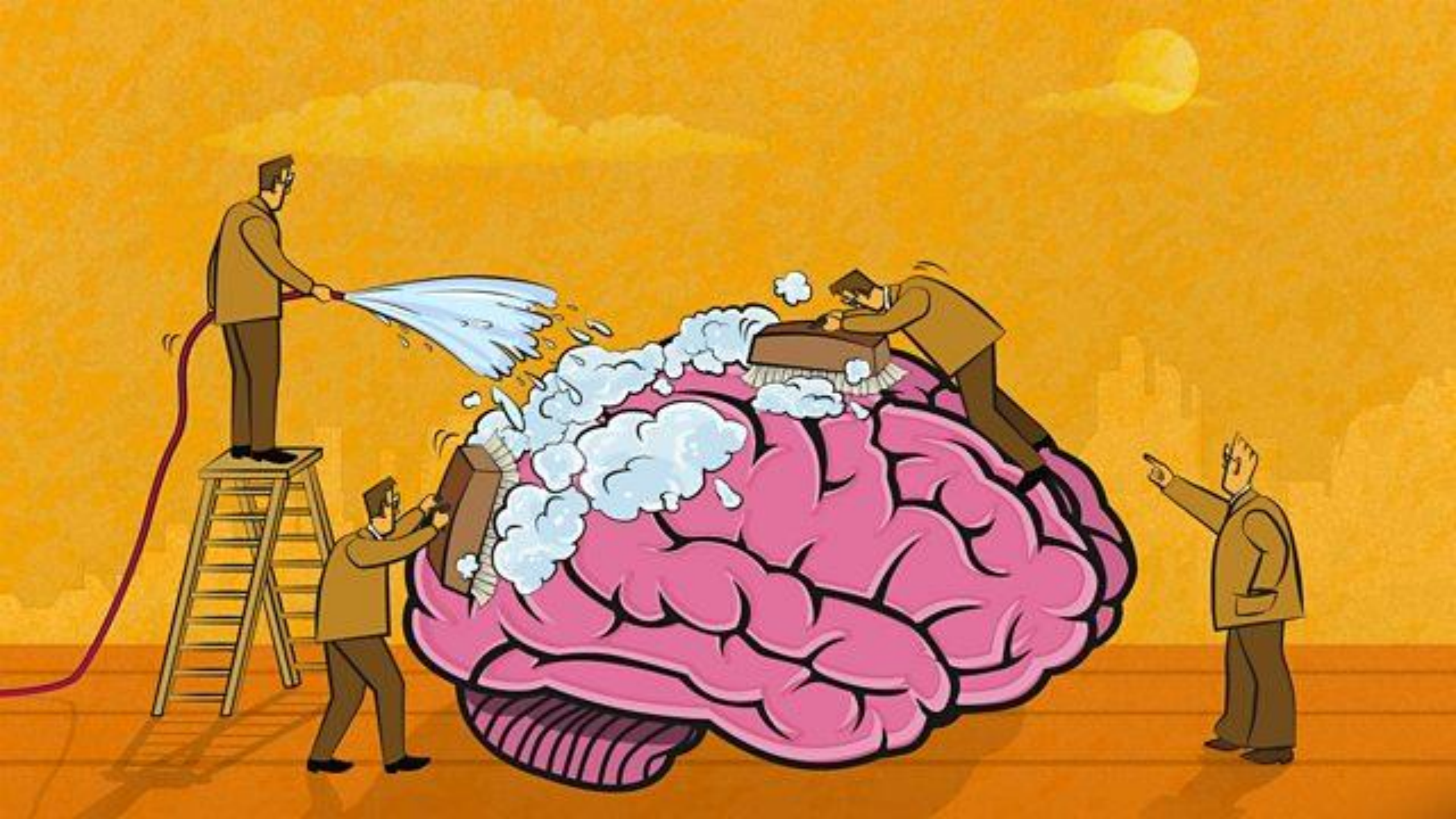


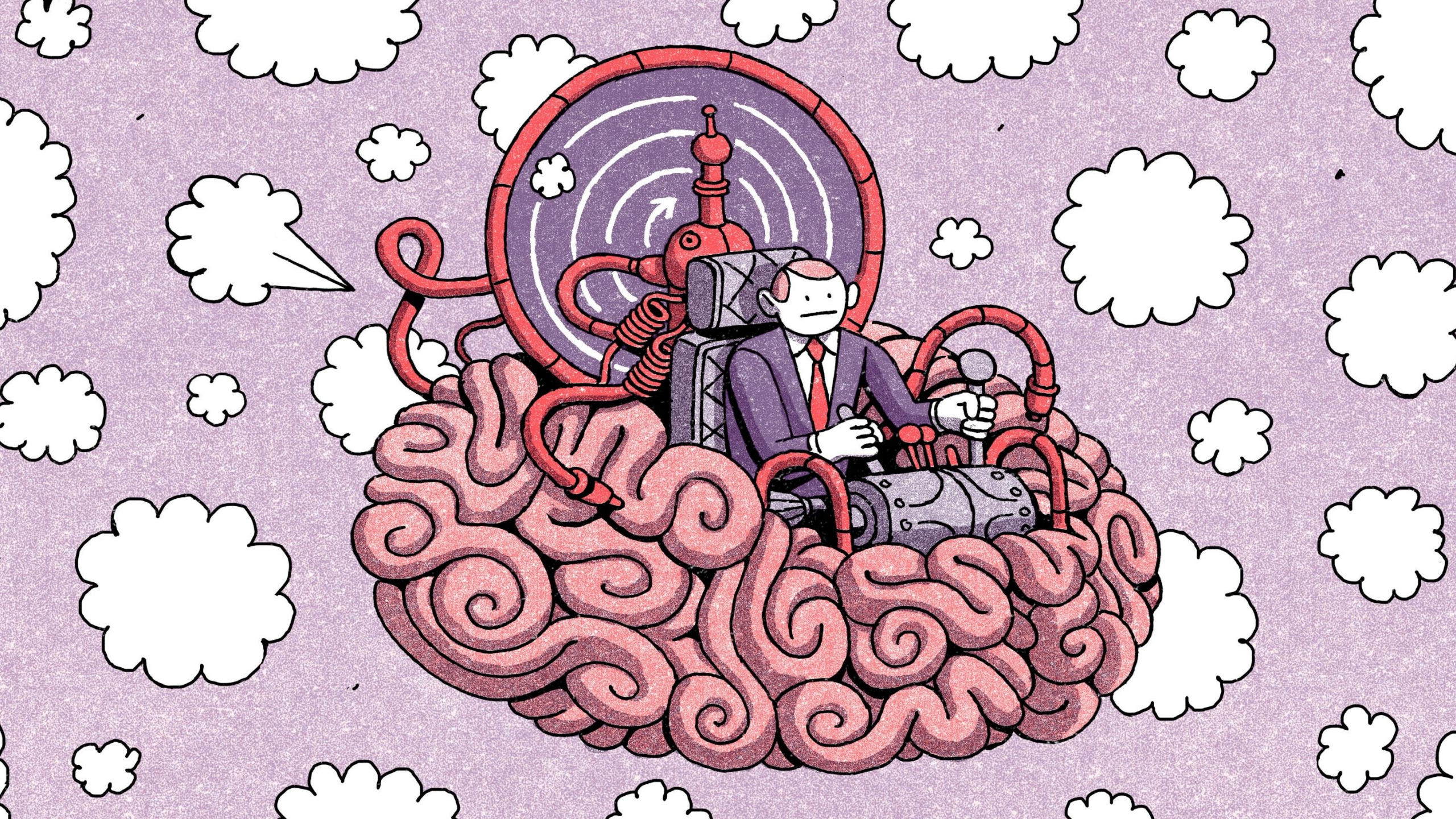










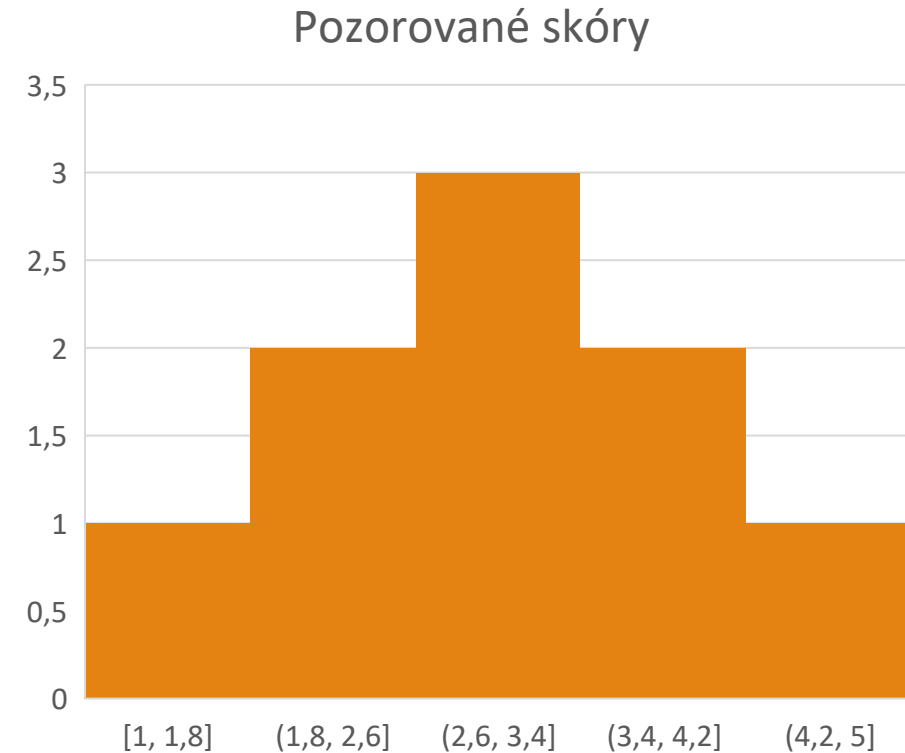


A close-up, high-contrast photograph of a deep red velvet curtain. The fabric is heavily draped and folded, creating a complex pattern of ridges and valleys. The lighting is dramatic, with bright highlights on the ridges and deep shadows in the folds, emphasizing the texture of the material.

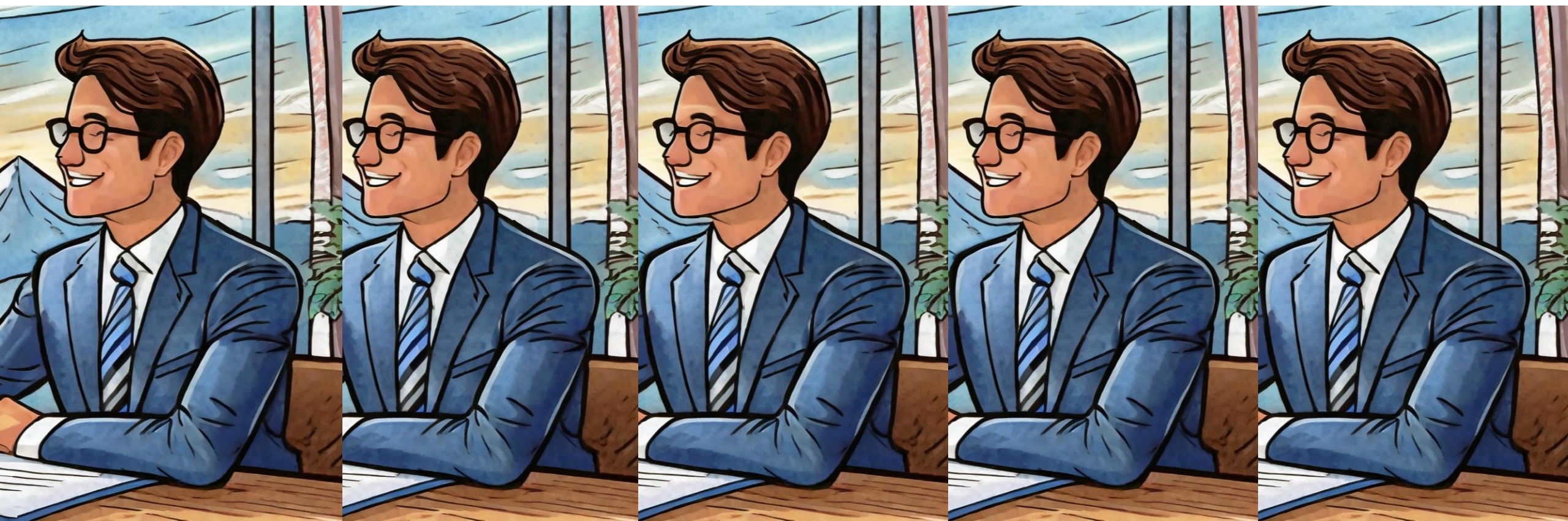
Directed by
DAVID LYNCH

Výsledek myšlenkového experimentu

Pozorovaný skór	Pravý skór	Chyba
3	3	0
3	3	0
3	3	0
4	3	1
4	3	1
2	3	-1
2	3	-1
1	3	-2
5	3	2

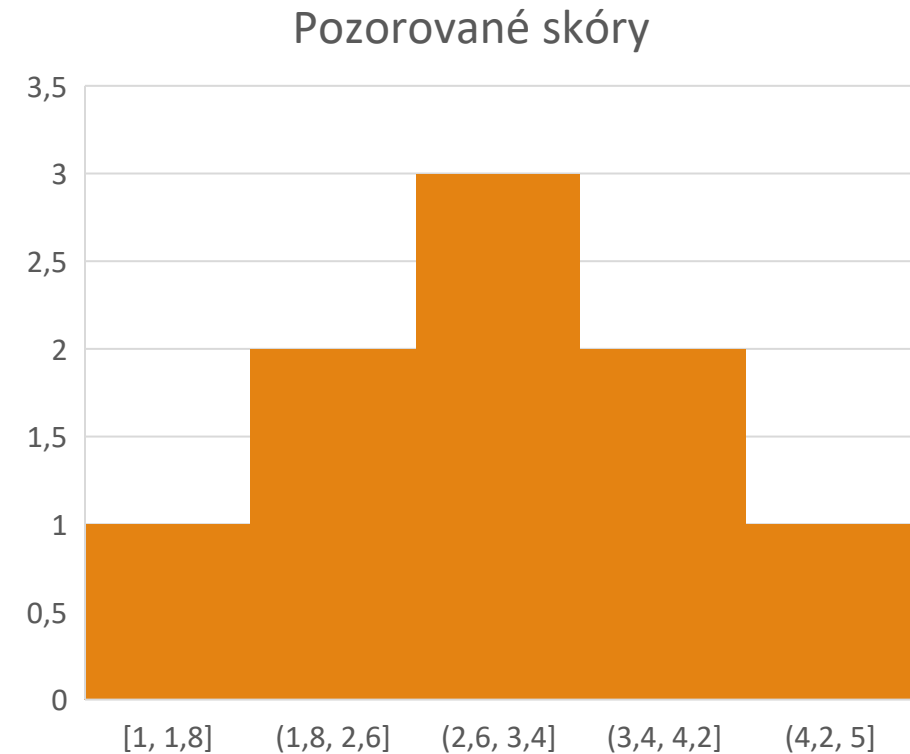


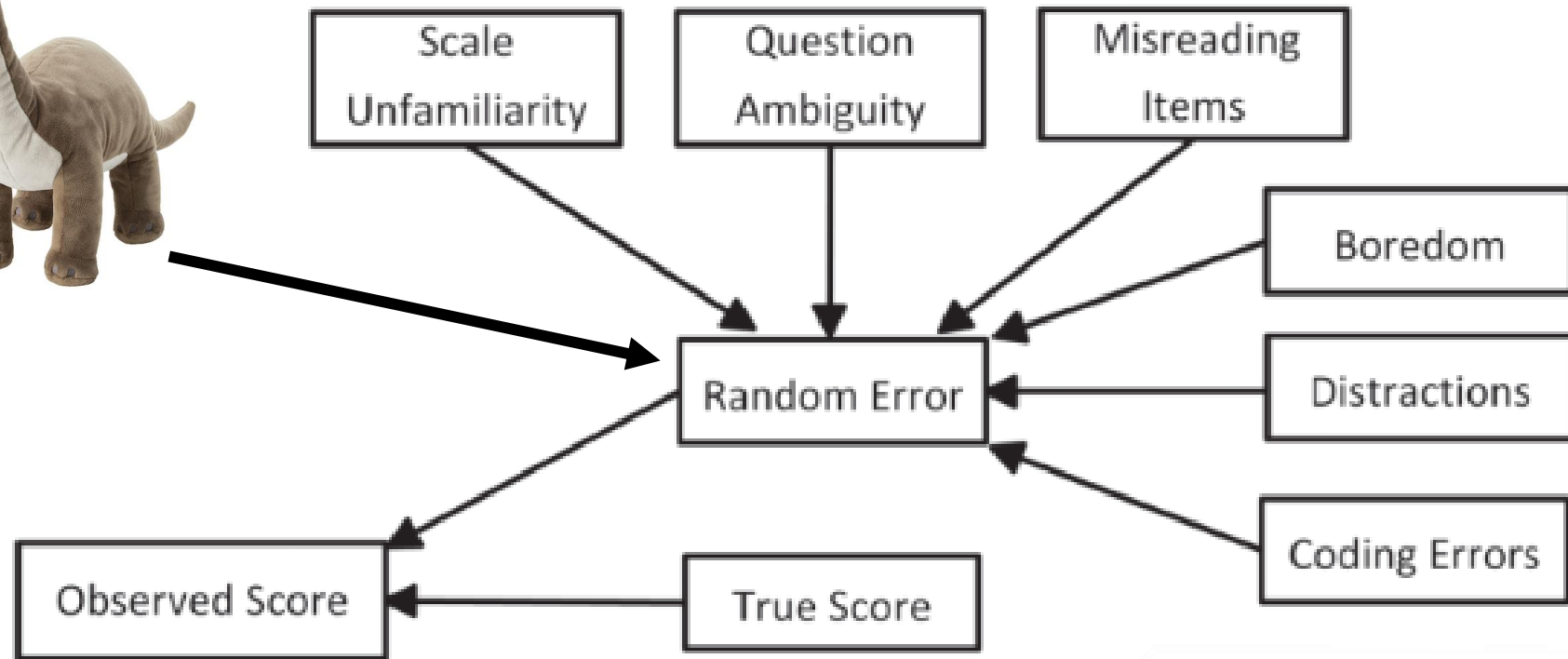
Alternativa?



Výsledek myšlenkového experimentu

Pozorovaný skór	Pravý skór	Chyba
3	3	0
3	3	0
3	3	0
4	3	1
4	3	1
2	3	-1
2	3	-1
1	3	-2
5	3	2





„Where does error come from? The conceptualization of the true score as an expected value is ill-defined. **For it is entirely unclear under what circumstances the replications mentioned in Lord and Novick’s brainwashing thought experiment should occur. The primary problem is that it is unclear where the random variation is supposed to come from.** This issue is usually circumvented in treatises on psychological measurement. These suggest that random error is due to unsystematic factors affecting the observations. For example, the **typical examples of unsystematic errors** are: Mr Brown had a **headache at the particular testing occasion**; Mr Brown **accidentally filled in ‘yes’**, while he intended to fill in ‘no’; Mr Brown **was distracted by the noise of schoolchildren playing nearby**, etc. However, identifying this, in itself reasonable, conceptualization of random error with the formal term indicated by E_i is circular.“ (Boorsboom, 2005, s. 36)

Jak to obejít v praxi?



OK, jak to teda ale udělat v reálném životě?

Jak to obejít v praxi?



OK, jak to teda ale udělat v reálném životě?

1. Přesun na úroveň populace (resp. vzorku).
2. Pokud chyba vyjadřuje nepřesnost, tak musí mít *opak* v přesnosti, right?
3. Vyjádřením míry přesnosti může být **reliabilita**.
4. Reliabilitu můžeme odhadnout z pozorovaných skóre několika lidí.
5. A potom vypočítat chybu měření z reliability.

A close-up shot of a man with short, light brown hair and blue eyes. He has a wide-eyed, questioning expression on his face. He is wearing a dark red jacket over a matching red shirt. The background is a solid, vibrant red with a subtle vertical crease or shadow down the center.

Is it the future or is it the past?



Teorémy CTT

1. $\tau_i = E(X_i)$ (na úrovni jednotlivce)
2. $e_i = X_i - \tau_i$ (na úrovni jednotlivce)
3. $X = T + E$ (na vzorku populace)

	Pozorovaný skór	Pravý skór	Chyba
Dale Cooper	120	130	-10
Albert Rosenfield	145	120	25
Harry Truman	95	110	-15

Rozšíříme vzorek...

	Pozorovaný skór	Pravý skór	Chyba
Dale Cooper	120	130	-10
Albert Rosenfield	145	120	25
Harry Truman	95	110	-15
Gordon Cole	85	100	-15
Hawk	115	90	25
Andy Brennan	70	80	-10
<i>M</i>	105	105	0
<i>SD</i>	27,02	18,71	19,49
<i>Rozptyl</i>	730	350	380

Teorémy CTT

1. $\tau_i = E(X_i)$ (na úrovni jednotlivce)
2. $E_i = X_i - \tau_i$ (na úrovni jednotlivce)
3. $X = T + E$ (na vzorku populace)
4. $\sigma_X^2 = \sigma_\tau^2 + \sigma_e^2$ (na vzorku populace)

Proč?

- **Chyba je nezávislá na měřeném** (jinak by nebyla náhodnou chybou).
- Chyba a pravý skór tak spolu nemají vůbec žádný vztah. To můžeme vyjádřit nulovou korelací, resp. kovariancí ($r_{\tau e} = 0$).
- **Jaká pravidla platí pro sčítání rozptylů proměnných?**

Minivýlet do algebry

Lze zobecnit na korelační/kovarianční matice.

Σ_{AB}	A	B
A	σ_A^2	cov_{AB}
B	cov_{AB}	σ_B^2

$$\sigma_{A+B}^2 = \sigma_A^2 + \sigma_B^2 + 2cov_{AB}$$

Minivýlet do algebry

Lze zobecnit na korelační/kovarianční matice.

Σ_{AB}	A	B
A	σ_A^2	cov_{AB}
B	cov_{AB}	σ_B^2

$$\sigma_{A+B}^2 = \sigma_A^2 + \sigma_B^2 + \mathbf{2}cov_{AB}$$

Σ_{TE}	T	E
T	σ_A^2	0
E	0	σ_B^2

$$\sigma_{T+E}^2 = \sigma_X^2 = \sigma_T^2 + \sigma_E^2 + \mathbf{2 * 0}$$



Klasická testová teorie

Je proto logické ptát se – jakou část mého měření tvoří chyba?

Jinými slovy, pokud: $\sigma_X^2 = \sigma_\tau^2 + \sigma_e^2$

Kolik je $\frac{\sigma_\tau^2}{\sigma_X^2}$ Jaký je podíl rozptylu pravého skóru v celkovém rozptylu?

Odpovědi se říká reliabilita.

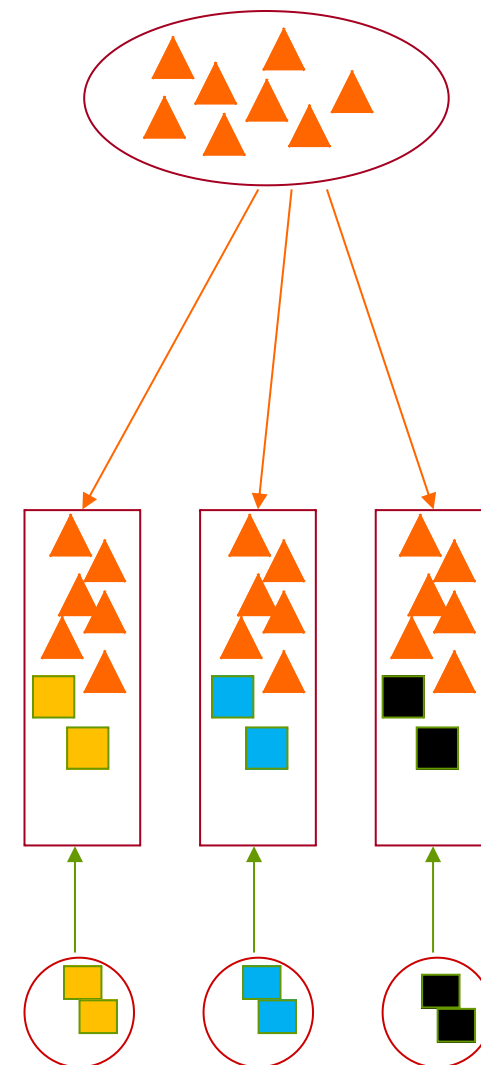
Reliabilita: metaforicky

Reliabilita: podíl společného a celkového rozptylu: $r_{xx'} = \frac{\text{▲}}{\text{▲} + \text{■} + \text{■} + \text{■}}$

Čím více „společné variability“ sdílí položky, tím vyšší je reliabilita testu (setu položek).

Více společného rozptylu → vyšší reliabilita.

Více chybového (specifického) rozptylu → nižší reliabilita.



Reliabilita: technicky

Reliabilita je definovaná jako podíl rozptylu pozorovaného skóre (manifestní proměnné) vysvětleného pravým skóre (latentní proměnnou):

$$r_{xx'} = (R^2) = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

Vysvětlený rozptyl je druhá mocnina korelace, tedy:

- $r_{xx'} = r_{x\tau}^2 = (R^2)$
- $\sqrt{r_{xx'}} = r_{x\tau} = (R)$
- Jinými slovy: reliabilita je umocněná korelace pravého a pozorovaného skóre.

OK. Ale jak tedy zjistíme reliabilitu?

	Pozorovaný skór	Pravý skór	Chyba
Dale Cooper	120	130	-10
Albert Rosenfield	145	120	25
Harry Truman	95	110	-15
Gordon Cole	85	100	-15
Hawk	115	90	25
Andy Brennan	70	80	-10
<i>M</i>	105	105	0
<i>SD</i>	27,02	18,71	19,49
<i>Rozptyl</i>	730	350	380

$$r_{xx'} = \frac{\sigma_{\tau}^2}{\sigma_x^2} = \frac{\sigma_{\tau}^2}{\sigma_{\tau}^2 + \sigma_e^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

Kde je háček?

	Pozorovaný skór	Pravý skór	Chyba
Dale Cooper	120	Neznáme	Neznáme
Albert Rosenfield	145	Neznáme	Neznáme
Harry Truman	95	Neznáme	Neznáme
Gordon Cole	85	Neznáme	Neznáme
Hawk	115	Neznáme	Neznáme
Andy Brennan	70	Neznáme	Neznáme
<i>M</i>	105	Neznáme	Neznáme
<i>SD</i>	27,02	Neznáme	Neznáme
<i>Rozptyl</i>	730	Neznáme	Neznáme

$$r_{xx'} = \frac{\sigma_{\tau}^2}{\sigma_x^2} = \frac{\sigma_{\tau}^2}{\sigma_{\tau}^2 + \sigma_e^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$



Jak to obejít v praxi?



OK, jak to teda ale udělat v reálném životě?

1. Přesun na úroveň populace (resp. vzorku).
2. Pokud chyba vyjadřuje nepřesnost, tak musí mít *opak* v přesnosti, right?
3. Vyjádřením míry přesnosti může být **reliabilita**.
4. Reliabilitu můžeme **odhadnout z pozorovaných skóre několika lidí**.
5. A potom **vypočítat chybu měření z reliability**.

Jak teda tu reliabilitu odhadneme?!

	Pozorovaný skór
Dale Cooper	120
Albert Rosenfield	145
Harry Truman	95
Gordon Cole	85
Hawk	115
Andy Brennan	70
<i>M</i>	105
<i>SD</i>	27,02
<i>Rozptyl</i>	730

OK, jak to teda ale udělat v reálném životě?

1. Přesun na úroveň populace (resp. vzorku).
2. Pokud chyba vyjadřuje nepřesnost, tak musí mít *opak* v přesnosti, right?
3. Vyjádřením míry přesnosti může být **reliabilita**.
4. Reliabilitu můžeme **odhadnout z pozorovaných skóre několika lidí**.
5. A potom **vypočítat chybu měření z reliability**.

Take home messages

1. CTT je hlavně teorie o chybě měření.
2. CTT je vystavěná na teorii chyb, ale nutně tak stojí na (v naší realitě) nefunkčním myšlenkovém experimentu.
3. Pro odhad chyby jednoho měření tak musíme jít na úroveň populace a spolehnout se na empirický odhad reliability celého testu.
4. Twin Peaks je must-see (doporučují 3/3 vašich vyučujících).

