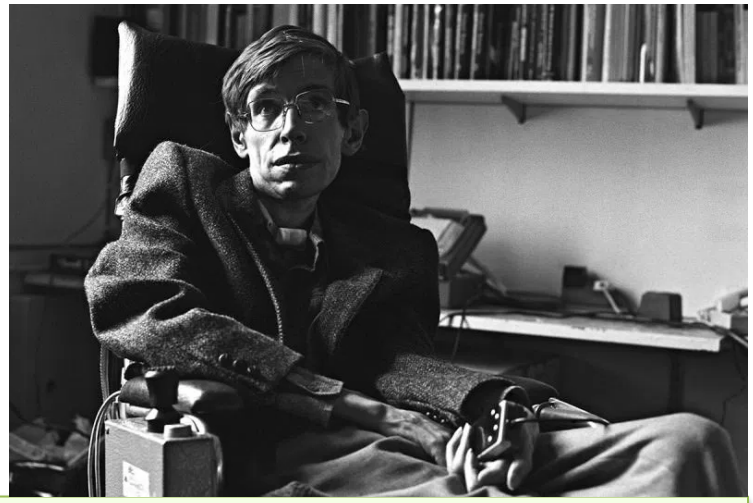
Implications éthiques et sociales de l'IA

Martin Gibert (IVADO, CRÉ) - UPOP 2018

"And, just like that, Facebook is giving us ads for used cars, optometrists, and couples counselling."

Les espoirs et les craintes de Stephen Hawking (1942-2018)



2014 « Je pense que le développement d'une intelligence artificielle complète pourrait **mettre fin à l'humanité**. Une fois que les humains auraient développé l'IA, celle-ci décollerait seule, et se redéfinirait de plus en plus vite. Les humains, limités par une lente évolution biologique, ne pourraient pas rivaliser et seraient dépassés. »

2016 « Nous ne pouvons pas prédire ce que nous pourrions réussir quand nos esprits seront amplifiés par l'IA. Peut-être qu'avec les outils de cette nouvelle **révolution technologique**, nous serons capables de défaire certains dégâts sur la nature de la précédente – l'industrialisation. »

Types d'IA et de problèmes.

- Types d'IA selon la fonction: calculer, classer, archiver, communiquer, diriger (voiture), fabriquer (robot), modéliser…
- IA faible ou forte.
- IA étroite ou générale.
- IA simple ou avancée.
- **Voiture autonome** = IA (faible) étroite et avancée
- **Super-intelligence** = IA (forte) générale et avancée.

- Problèmes **épistémiques** (impénétrabilité) ou **éthiques** (prise de décision)
- IA trop stupides (erreur, accident) ou trop **intelligentes** (perte de contrôle)
- Court terme ou long **terme**
- Nouveaux problème (SALA) ou anciens **démultipliés** (propagande)

Un peu de méthode

Paula Boddington: « L'éthique existe parce que le monde n'est pas parfait et que nous pensons que nous pouvons l'améliorer si nous essayons suffisamment. (…) Mais le monde est imparfait *d'une manière très compliquée*. Il est souvent difficile de se figurer ce qui *précisément* va mal, et a fortiori ce qu'on devrait faire. »

- L'éthique appliquée à l'IA est un domaine **diversifié** et **nouveau**.
- Attention à la **panique morale** et à la ***hype*.**
- Attention à ne pas négliger l'aspect **anxiogène** (et le manque de littératie numérique).
- Attention à l'**imagination** (ex. stéréotypes sur le futur).

Quatre dystopies contemporaines

1. **Anéantir l'humain**: une super-intelligence hostile qui prend le pouvoir.
2. **Remplacer l'humain**: une aristocratie d'élites améliorées qui règne sur des chômeurs drogués.
3. **Surveiller et punir l'humain**: un régime totalitaire néofasciste qui s'appuie sur l'IA.
4. **Manipuler l'humain**: une dictature douce où l'IA devance nos désirs.

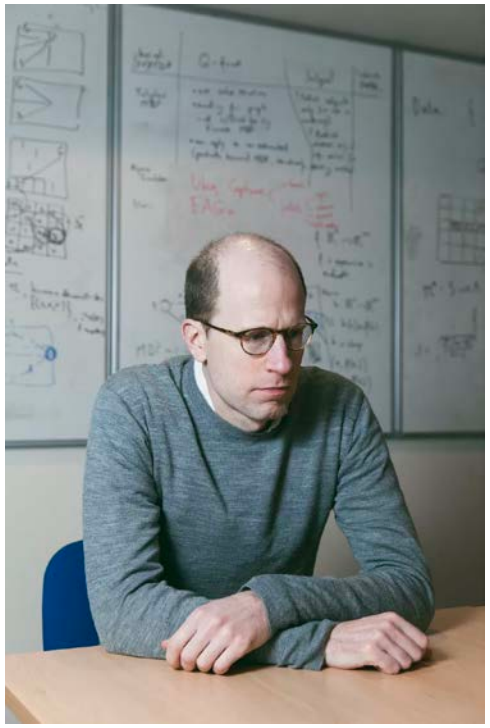
≠ Servir l'humain.

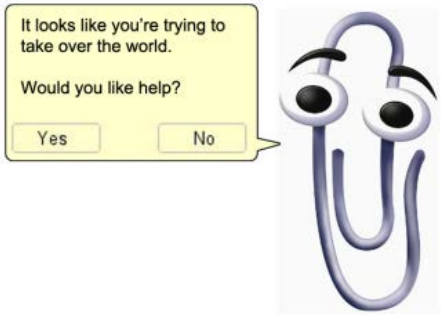
1- Anéantir

"I think I was only invited for one reason."

Bug

Une super-intelligence hostile



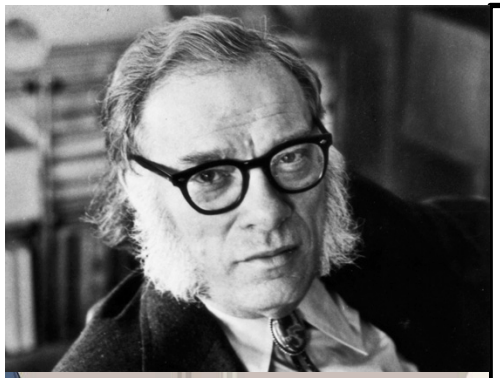


Nick Bostrom (*Future of Humanity Institute*)

- ex du **Paperclip maximizer/roi Midas.**

« Nous devons penser à l'intelligence comme un **processus d'optimisation**, un processus qui guide le futur dans un certain jeu de configurations. Une super intelligence est un processus d'optimisation très fort. Elle est très douée pour utiliser les moyens disponibles pour atteindre un état dans lequel son but est réalisé. Il n'y a pas de connexion nécessaire entre le fait d'être très intelligent en ce sens, et avoir un d'objectif que nous, humains, trouverions utile ou significatif. » TED Talk 2017

Les lois d’Asimov (1942)

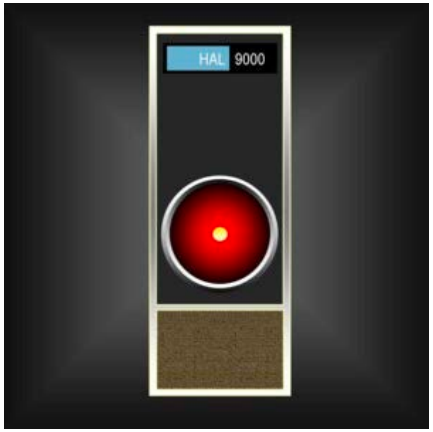


1. Un robot ne peut **porter atteinte à un être humain**, ni, en restant passif, permettre qu'un être humain soit exposé au danger ;
2. Un robot doit **obéir aux ordres** qui lui sont donnés par un être humain, sauf si de tels ordres entrent en conflit avec la première loi ;
3. Un robot doit **protéger son existence** tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.

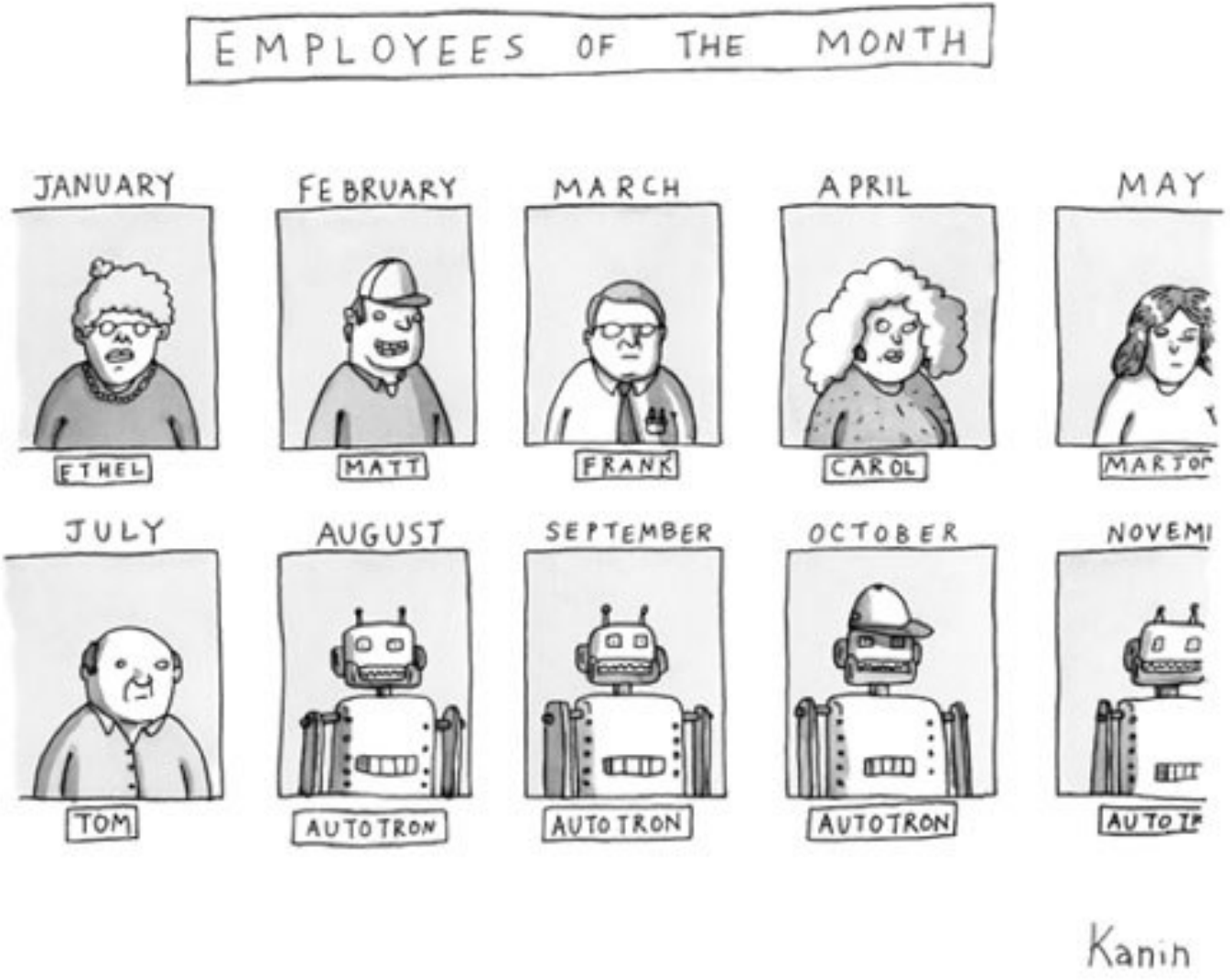
- Loi Zéro (1950) : Un robot ne peut pas **faire de mal à l'humanité**, ni, par son inaction, permettre que l'humanité soit blessée.
- (= Principe responsabilité)

Vers la fin du monde?

- **Risque existentiel**: « Ce sont des menaces qui pourraient causer notre extinction ou détruire le potentiel de la vie intelligente sur Terre. » (Bostrom 2002)
- Faut-il en parler?
- On peut toujours éteindre l'ordi?
- « Nous ne devrions pas faire confiance à notre capacité à garder éternellement un génie super intelligent prisonnier dans sa lampe.» (Bostrom 2017)
- **Réponse**: Développer une IA qui partage fondamentalement nos valeurs – et plus vite qu'une super-intelligence.

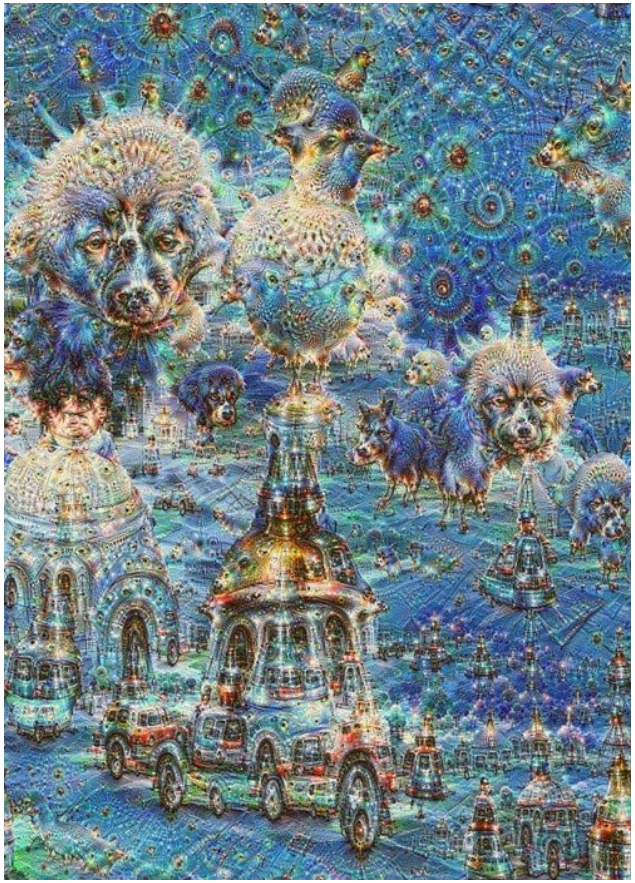


2- Remplacer

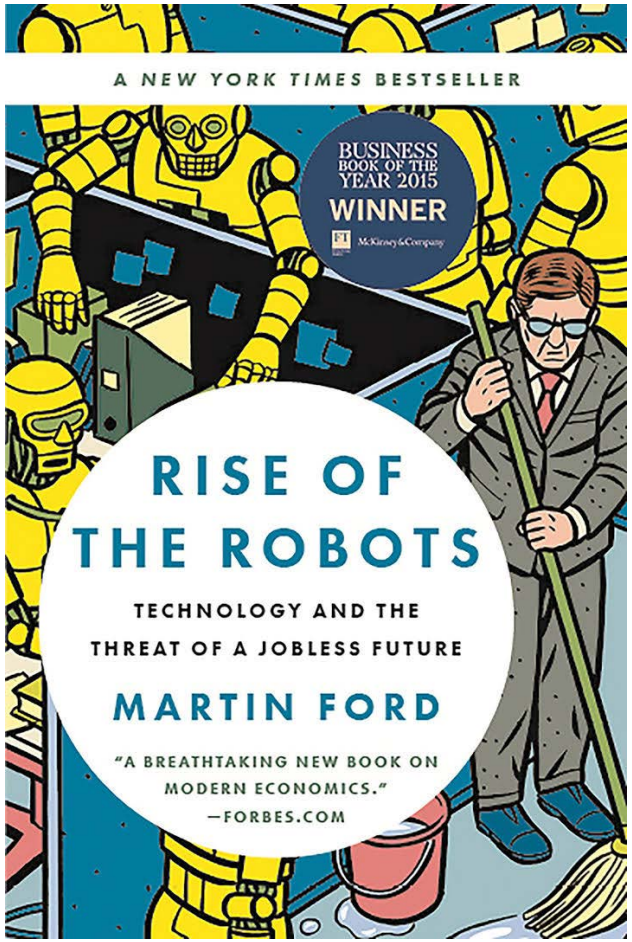

EMPLOYEES OF THE MONTH
JANUARY
ETHEL
FEBRUARY
MATT
MARCH
FRANK
APRIL
CAROL
MAY
JULY
TOM
AUGUST
AUTOTRON
SEPTEMBER
AUTOTRON
OCTOBER
AUTOTRON
Kanin

Remplacer quoi?

- Seule l'IA: *trading* à haute fréquence.
- L'apprentissage profond remplace bien les **tâches cognitives « intuitives »** ou système 1: conduite, reconnaissance d'image…
- Revalorisation de la **créativité**? (parce que l'IA n'est pas meilleure que ses données).
- Revalorisation du **soin**? (parce que l'IA simule mal l'empathie?)
- Attention à l'exemption de **responsabilité**: SALA, algo en justice.

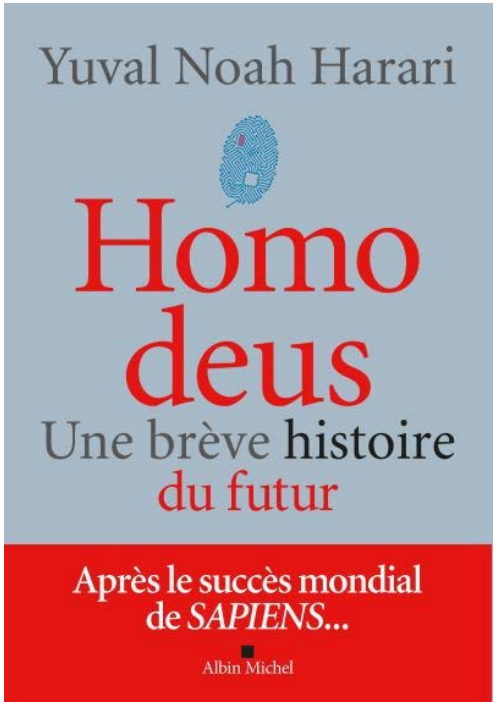


Remplacer qui?



- **Martin Ford** (2015) : contrairement aux révolutions industrielles précédentes peu de jobs seront créés par l'IA.
- L'IA remplacera aussi les travailleurs qualifiés (radiologistes, juristes, software designer)
- (Frey & Osborne 2013) **47%** des emplois US seraient automatisables (à 70%); (OCDE 2016) **9%** pour les 21 pays de l'OCDE.
- Il n'empêche que l'automatisation (et les tracteurs) crée de la richesse.
- Problème de **justice distributive**.

Vers une aristocratie d'élites améliorés?



- « *La manne technologique à venir permettra probablement de nourrir et de soutenir les gens, même sans aucun effort de leur part. Mais qu'est-ce qui les gardera occupés et contents? Une réponse pourrait être des médicaments et des jeux informatiques.* »
- **Monopoles**: Google contrôle 40% de la publicité digitale US, Facebook 20%.
- Nouvelles formes d'**inégalités** (biologiques)?

« Toutes les prédictions qui parsèment ce livre ne sont rien de plus qu'une tentative pour aborder les dilemmes d'aujourd'hui et une invitation à changer le cours de l'avenir. » Y. Harari

3- Surveiller et punir (1993 – 2015)

"On the Internet, nobody knows you're a dog."

"Remember when, on the Internet, nobody knew who you were?"

Discriminations algorithmiques

- *Machine bias* (ProPublica 2016)
- COMPAS = algorithme d'évaluation des risques de récidive vendu « pour réduire le taux d'incarcération »
- À partir de 137 questions: job, études, famille…mais pas la race.
- Faux positifs favorables aux blancs et faux négatifs défavorables aux noirs.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Des données déjà biaisées

(Résultat de recherche pour « CEO » sur Google image).

IA, justice et diversité

Cathy O'Neil, *Weapons of Math destruction*:

« Les algorithmes ne rendent pas les choses équitables si on les applique aveuglément, avec négligence. Ils n'instaurent pas l'équité. Ils reproduisent nos pratiques du passé, nos habitudes. Ils automatisent le statu quo. »

Timnit Gebru, *Laboratoire d'IA de Stanford*:

« Si nous n'avons pas de diversité dans notre groupe de chercheurs, nous n'aborderons pas les problèmes auxquels sont confrontés la majorité des gens dans le monde. »

Un droit à l'intériorité?

- L'IA peut percevoir mieux que l'humain.
- Ex. *Entretient d'embauche, détecteur de mensonge, app. de dating.*

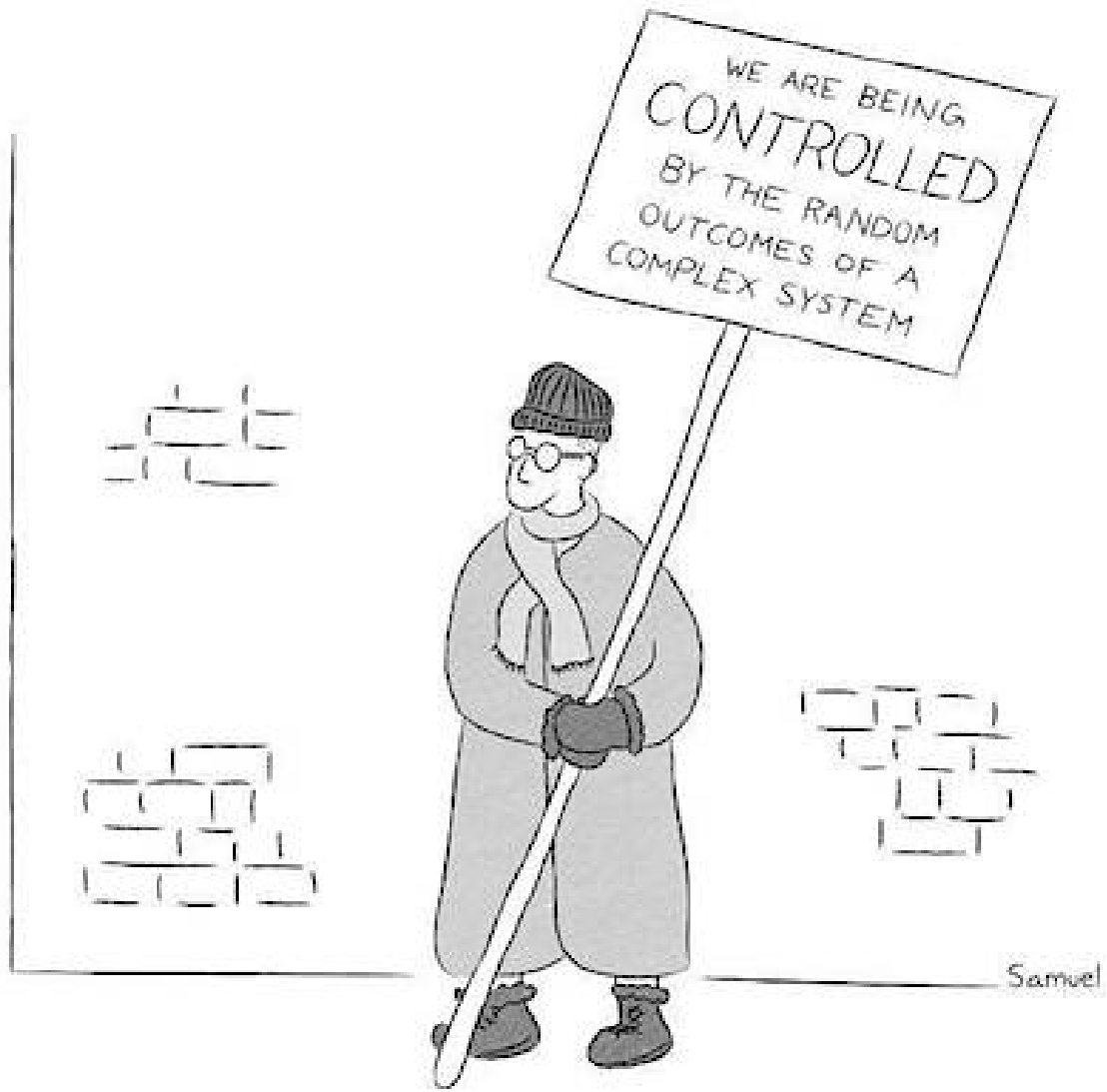
Jocelyn Maclure (*commission de l'éthique en sc. et technologie*)**:** « On doit garder le contrôle sur ce qu'on choisit de dévoiler par rapport à notre vie intérieure, à notre vie subjective, à notre vie intime. Le même questionnement se pose en neuro-éthique pour les technologies qui permettent d'étudier le cerveau et d'avoir accès à des informations relatives à la vie cérébrale des personnes.»

Vers une dictature digitale?

- *Ex. reconnaissance faciale, données biométriques.*
- Un officiel chinois à Davos: « À l'ouest, vous avez la démocratie, nous avons les média sociaux. »
- Qui devrait posséder les data? Comment les réguler ?
- **Vie privé** : désir d'intimité vs culture de l'exhibition (FB, selfie)
- Liberté vs **sécurité/santé**

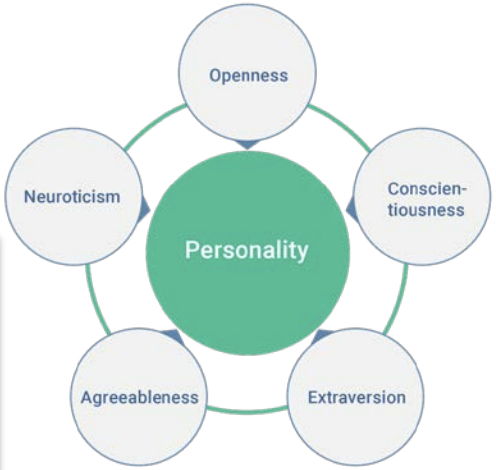
"If you're not planning to break the law, why should you care?"

4- Manipuler


WE ARE BEING
CONTROLLED
BY THE RANDOM
OUTCOMES OF A
COMPLEX SYSTEM
Samuel

L’affaire Cambridge Analytica

- = firme de conseil stratégique crée en 2013 spécialisée dans le profilage et le comportement électoral.
- 270 000 personnes font le Quiz, les données de 80 millions d’amis sont siphonnées.
- = violation du droit à la protection des données personnelles.
- Corrélations *likes* / traits de personnalité.
- = ciblage des consommateurs/électeurs.



Christopher Wylie: « Elles [FB et Google] pourraient dire par exemple: attention, ceci est une publicité, vous avez été visé et voilà qui paie pour ça. »

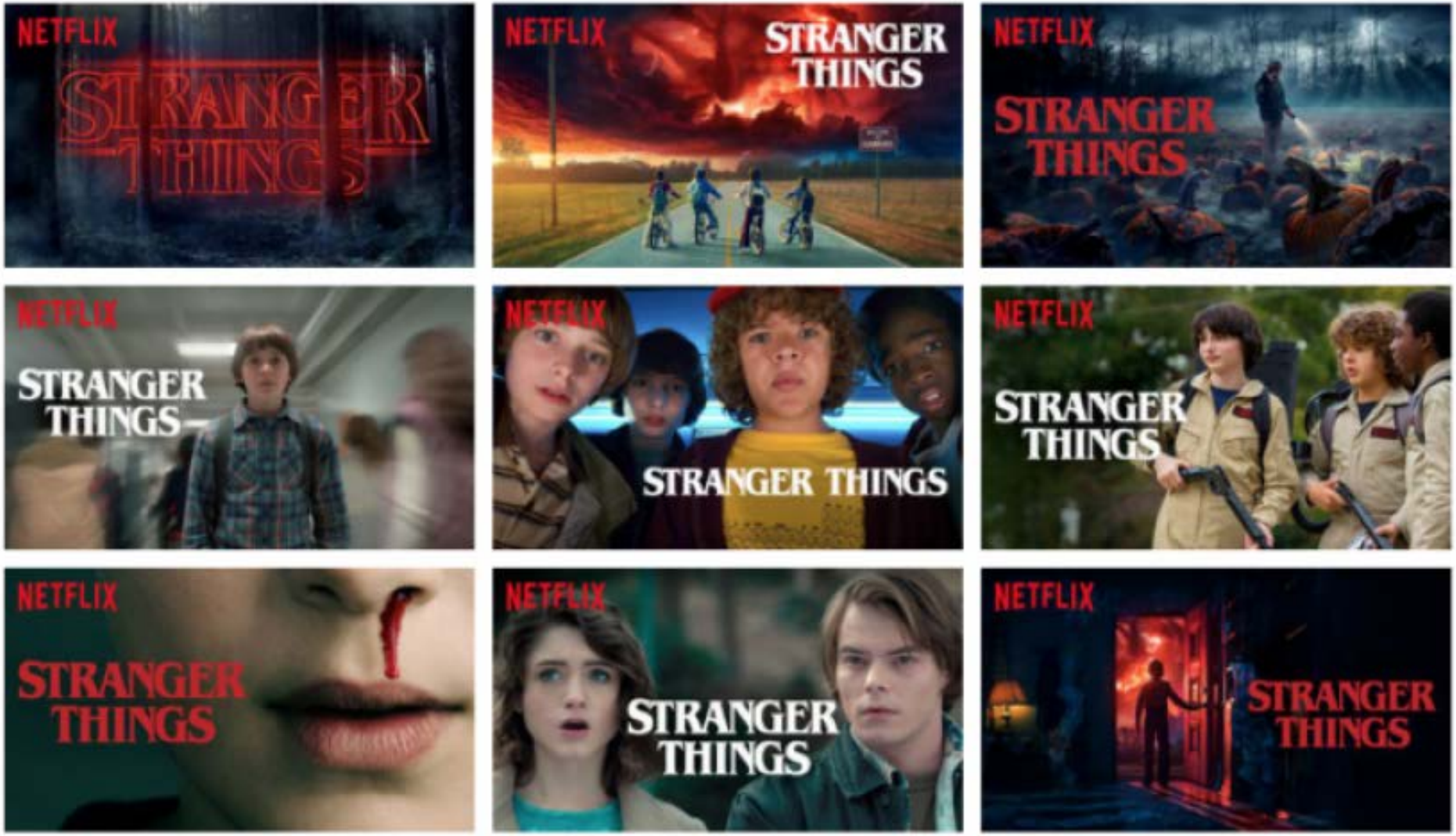
Le temps de cerveau disponible

Patrick Le Lay, PDG de TF1 (2004)

« Il y a beaucoup de façons de parler de télévision. Mais dans une perspective *business*, soyons réaliste: à la base, le métier de TF1, c'est d'aider Coca Cola, par exemple, à vendre son produit. Or pour qu'un message publicitaire soit perçu, il faut que le cerveau du téléspectateur soit disponible. Nos émissions ont pour vocation de le rendre disponible: c'est-à-dire de le divertir, de le détendre pour le préparer entre deux messages. Ce que nous vendons à Coca Cola, c'est du **temps de cerveau** humain disponible. »

Reed Hastings, CEO de Netflix (2017).

«Quand vous regardez une série sur Netflix et que vous en devenez accro, vous veillez tard le soir. À la marge, nous sommes en concurrence avec le **sommeil**. »


NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS
NETFLIX
STRANGER THINGS

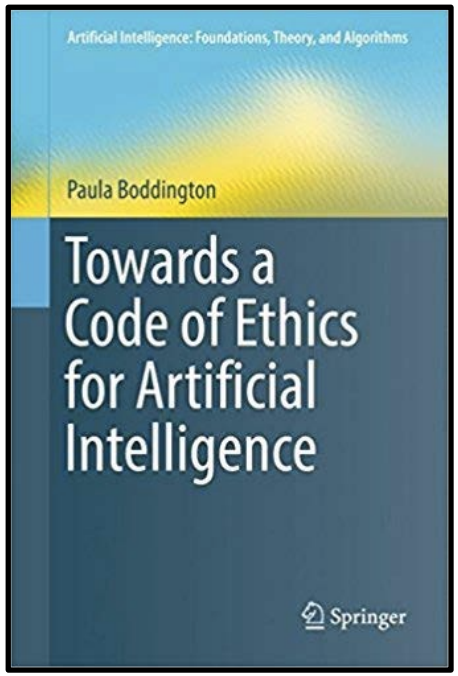
Comment « hijacker » le cerveau humain?

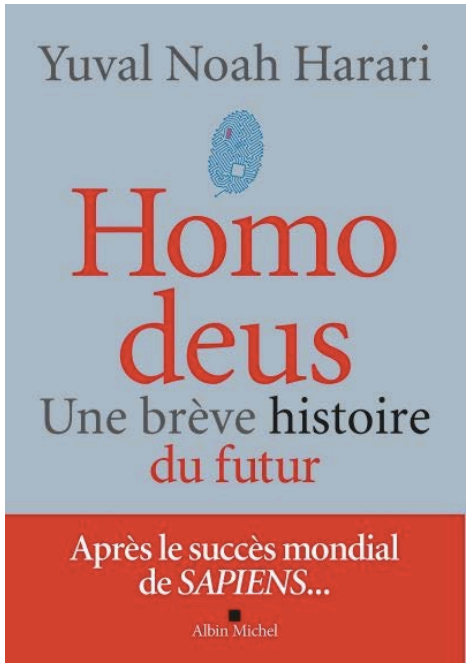
Tristan Harris : « Une fois que vous avez compris comment appuyer sur les boutons des gens, vous pouvez en jouer comme du piano. »

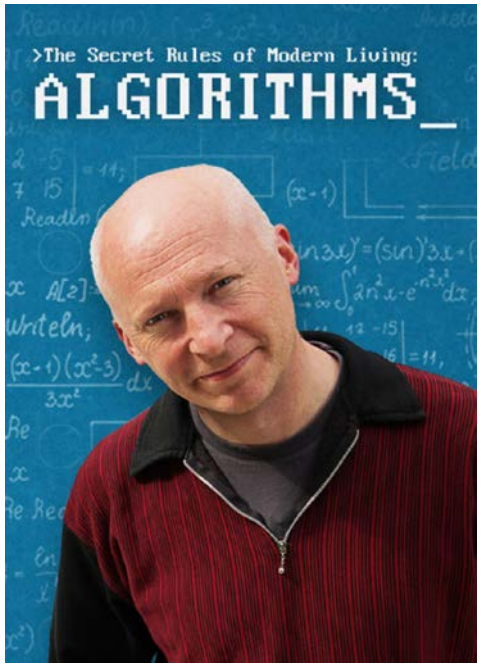
- Contrôlez le **menu** et vous contrôlerez les choix.
- Mettez une **machine à sous** dans la poche d'un milliard de personnes (notifications, scrolling, swipe sur Tinder).
- **FOMO** (fear of missing out): facebook, app de rencontre…
- Les gens recherchent **l'approbation sociale** : tag, nouvelle photo de profil.
- Utilisez des News feed **sans fin** et l'*autoplay* (Youtube).

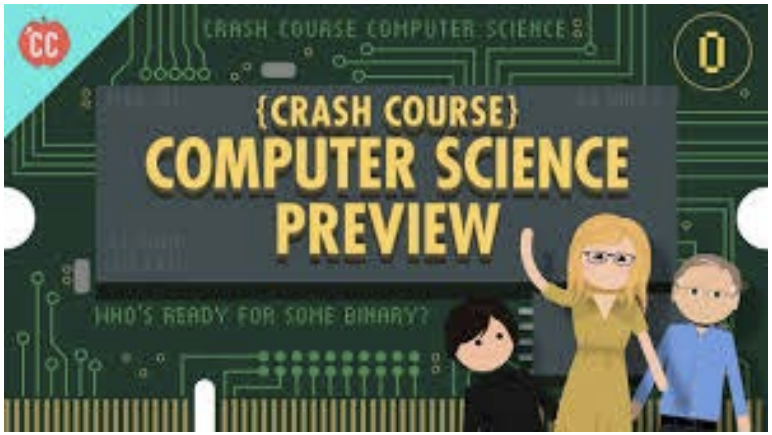
Vers une dictature douce?

- **Zeynep Tufekci** (TED talk): « Nous avons besoin d'une économie digitale où nos données et notre attention ne sont pas à vendre au plus offrant démagogue ou dictateur. »
- Mais il ne faut pas rejeter en bloc les médias


Artificial Intelligence: Foundations, Theory, and Algorithms
Paula Boddington
Towards a Code of Ethics for Artificial Intelligence
Springer


Yuval Noah Harari
Homo deus
Une brève histoire du futur
Après le succès mondial de SAPIENS...
Albin Michel


>The Secret Rules of Modern Living:
ALGORITHMS_


CRASH COURSE COMPUTER SCIENCE
{CRASH COURSE}
COMPUTER SCIENCE PREVIEW
WHO'S READY FOR SOME BINARY?

Merci