Les limites de la rationalité en intelligence artificielle

Charles Bodon

Les limites de la rationalité en intelligence artificielle

Charles Bodon

Université Paris 1 Panthéon-Sorbonne

bodonbruzel@gmail.com

Communication dans le cadre du cours « Intelligence artificielle : enjeux et applications » de M. Vladimir Atlani, Sciences Po. Paris

Résumé

L'intelligence artificielle (IA) suscite de nombreux fantasmes parmi les chercheurs en sciences cognitives d'Amérique du Nord et dont certains perdurent encore aujourd'hui dans les médias et le débat public. Souvent présentée comme une technologie révolutionnaire, ce qu'elle est en pratique, notamment dans ses techniques les plus récentes, elle rencontre pourtant encore de nos jours des limites théoriques (parfois indépassables) identifiées dès le 20e siècle. En effet, que ce soit dans la compréhension du contexte, l'interprétation de propositions sémantiquement ambiguës, ou tout simplement dans son traitement de l'information, la rationalité de l'IA demeure bornée à celle d'une approche logico-mathématique du monde. Il y a ainsi de nombreux abus de langage dans le débat public lorsque l'on évoque l'IA et qui conduisent à développer des croyances quant à sa réalité. On dira ainsi d'une machine qu'elle est « plus intelligente » que l'être humain (alors qu'elle n'est tout simplement pas intelligente), qu'elle « comprend » plus ou moins bien certains problèmes (alors qu'elle ne comprend pas des choses au même titre qu'un humain), ou encore qu'elle est un « système autonome » (alors qu'elle est rigoureusement déterminée dans toutes ses actions). Or, s'exprimer ainsi c'est méconnaître les limites réelles de l'IA qu'elle doit à sa nature mécanique et c'est poursuivre et prolonger les fantasmes que cette technologie alimente depuis les années 50. L'enjeu de notre présentation ne sera donc pas de dire que l'IA est inefficace (elle l'est bien évidemment dans de nombreux cas), mais plutôt d'amener un éclairage quant à certaines de ses limites et dissiper les malentendus quant à ses possibilités.

Objectifs

Désensorceler les représentations que l'on a de l'IA.

- L'IA n'est pas capable de tout. Elle peut très peu par rapport à l'humain.

Désambiguïser certaines expressions du langage courant.

- « La machine est plus intelligente que l'humain ».
- « La machine "comprend", "pense", "calcule" ».

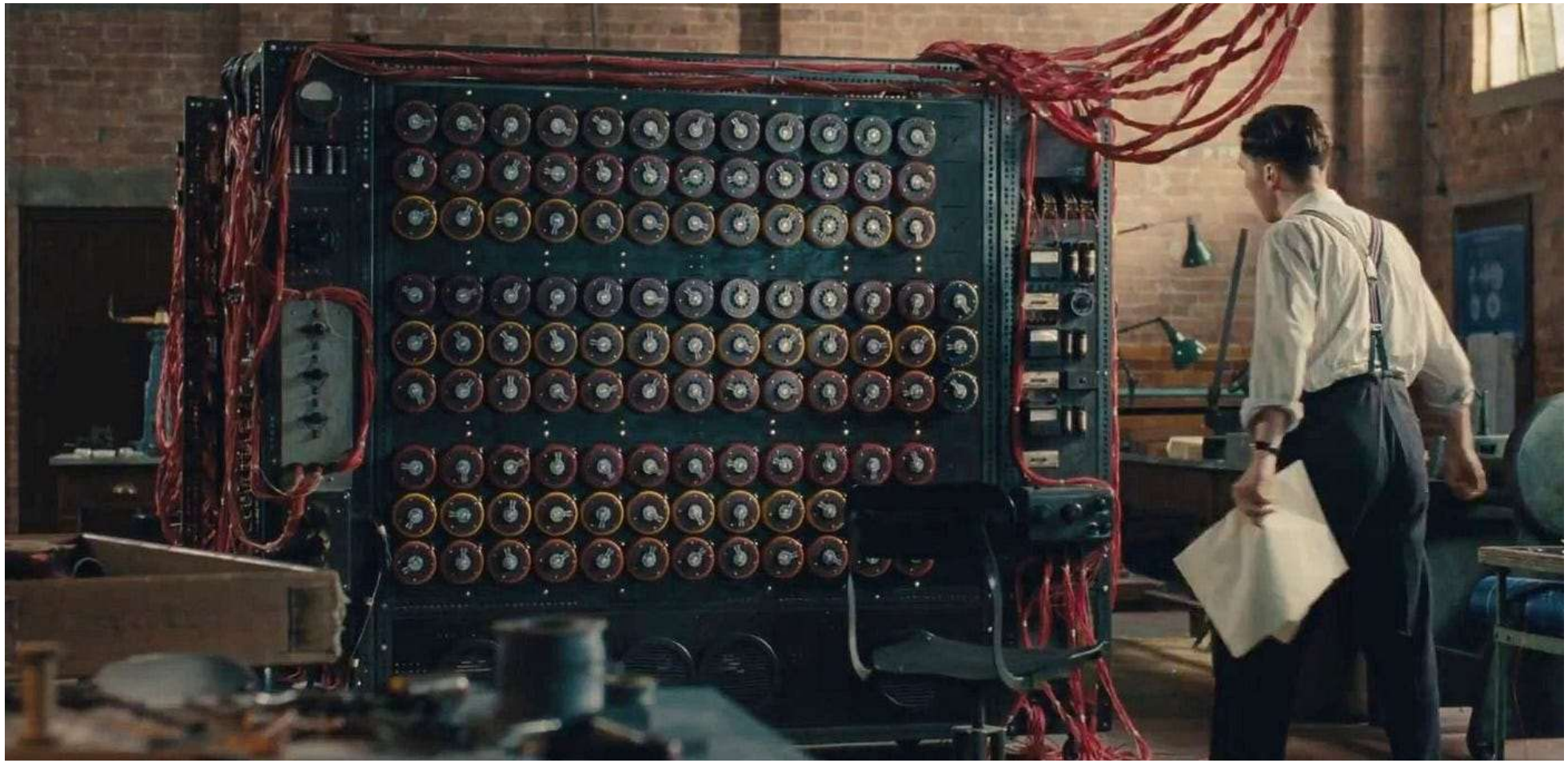
Mettre en perspective nos proximités et différences avec l'IA.

- L'erreur est humaine… et la machine l'imite.

Plan de la communication

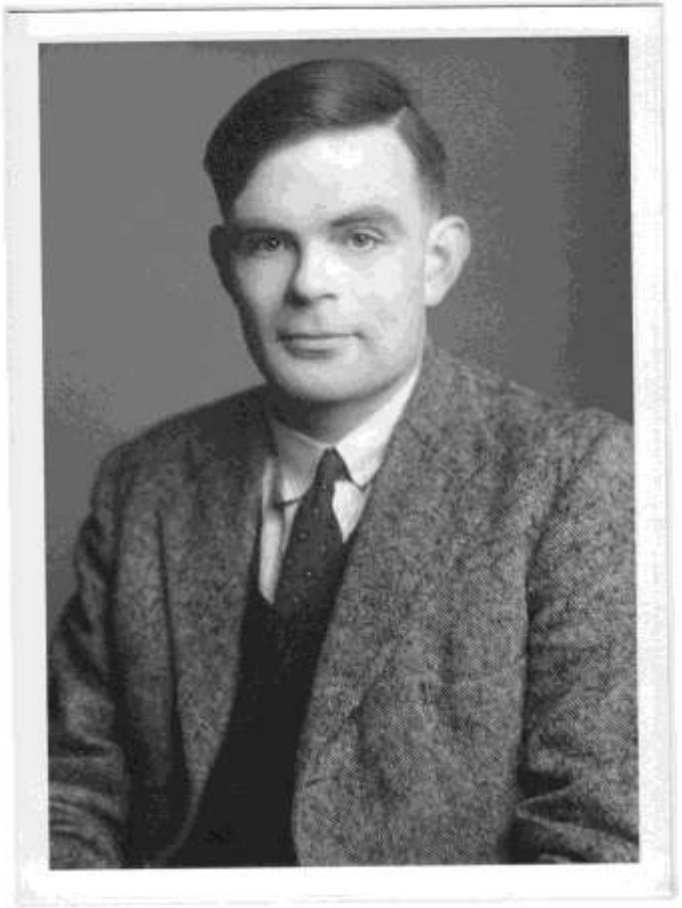
1. Principes de l'IA chez Turing.
2. Débats philosophiques autour des limites de l'IA.

1. Principes de l'IA chez Turing



The Imitation Game (2014)

La machine de Turing (1936)



Alan Turing (1912 – 1954)

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]

Trois objectifs théoriques

1. Donner une définition rigoureusement logico-mathématique de la **calculabilité**.
2. Donner une définition de ce qu'est une **procédure effective** (algorithme).
3. Donner une application de ces définitions au **problème de la décision** (*Entscheindungsproblem*).

Que fait le mathématicien quand il calcule ?

1. Lit et écrit des **nombres**.
2. Utilise des **symboles** (1, 2, +, =, x, n, $\sqrt{}$, $\forall$, etc.).
3. Suit des **règles** (addition, soustraction, multiplication, division).
4. Utilise sa **mémoire**.
5. Utilise différents **systèmes**.

La machine de Turing, une « machine de papier »

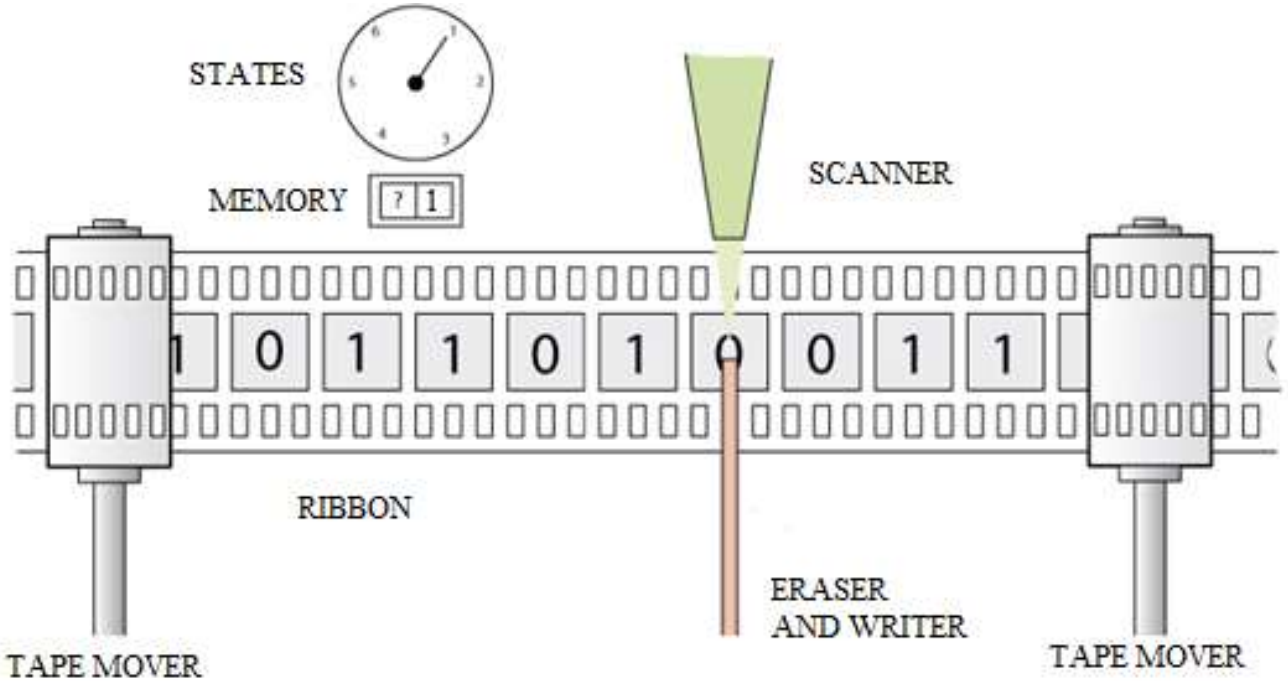


Figure 1. Représentation imagée d'une machine de Turing
(Hao Wang, *Games, Logic, and Computers*, Scientific American, Volume 213, Number 5, November 1965, pages 98-106)

1. Un ruban divisé en plusieurs cases et contient un alphabet.
2. Un scanner qui lit et écrit.
3. Une mémoire qui enregistre.
4. Une table d'instruction.
5. Différents états.

Un ordinateur théorique

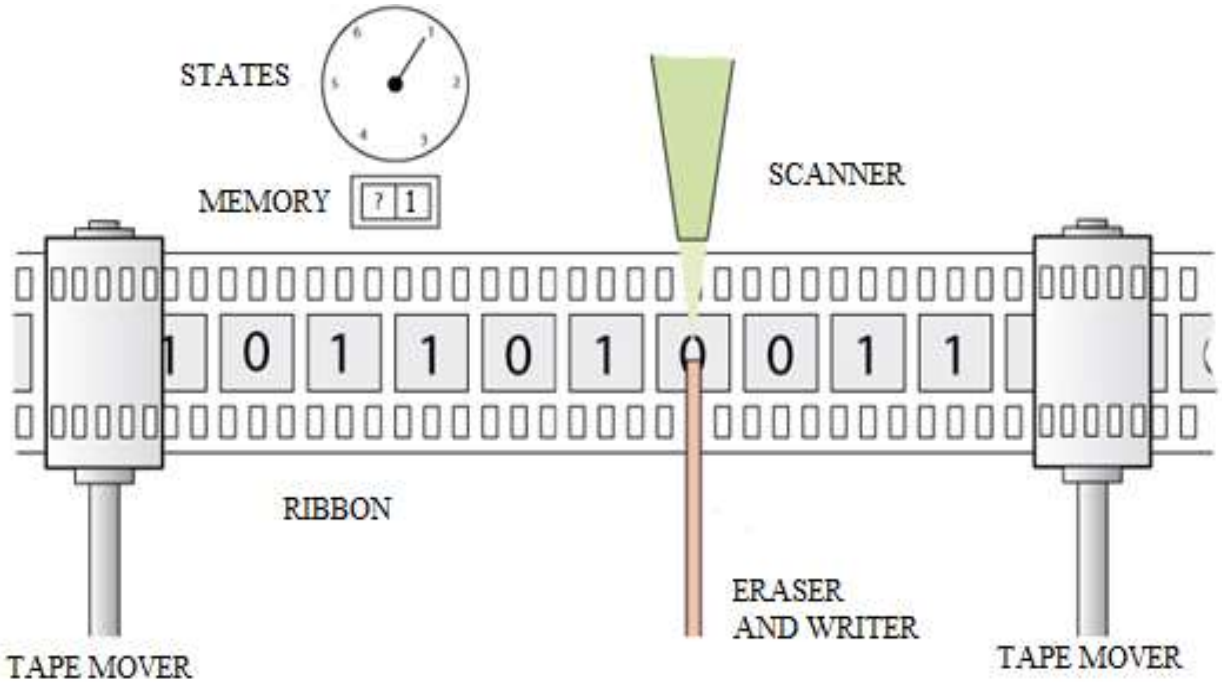


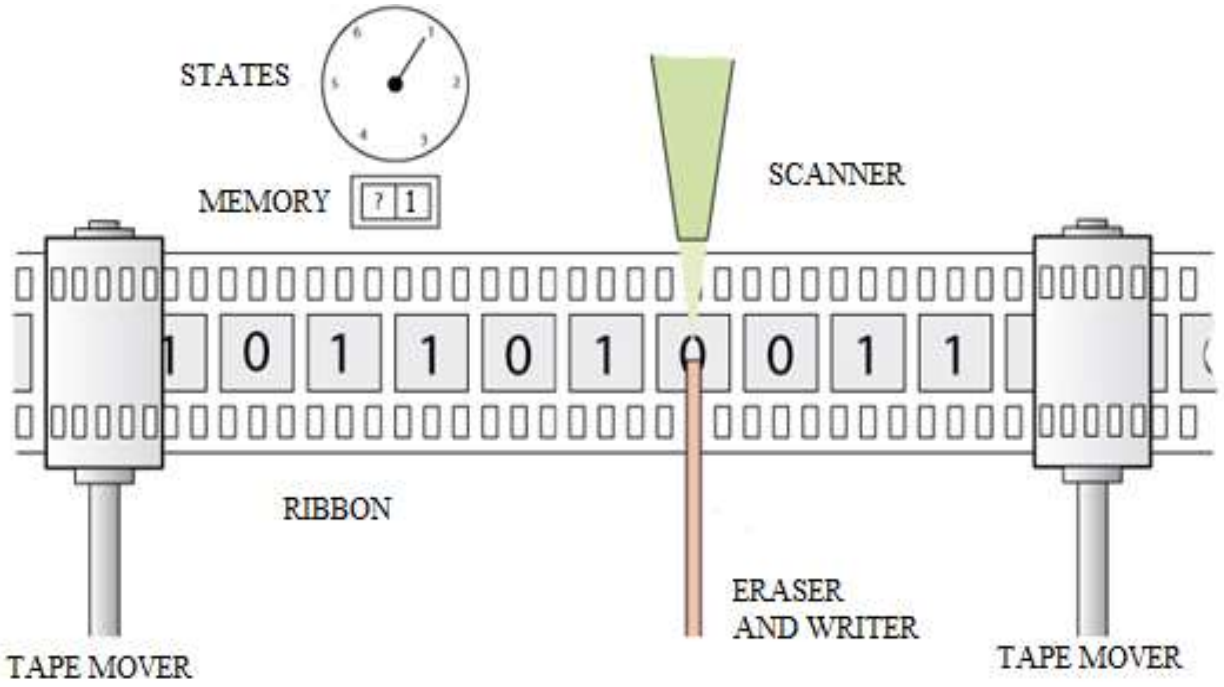
Table d'instruction

État 1 = Si tu lis « 0 » alors va à droite, écris « 0 », puis passe en « État 2 ».

État 2 = Si tu lis 0 alors va à droite et écris « 1 », puis passe en « État 3 ».

État 3 = Halte.

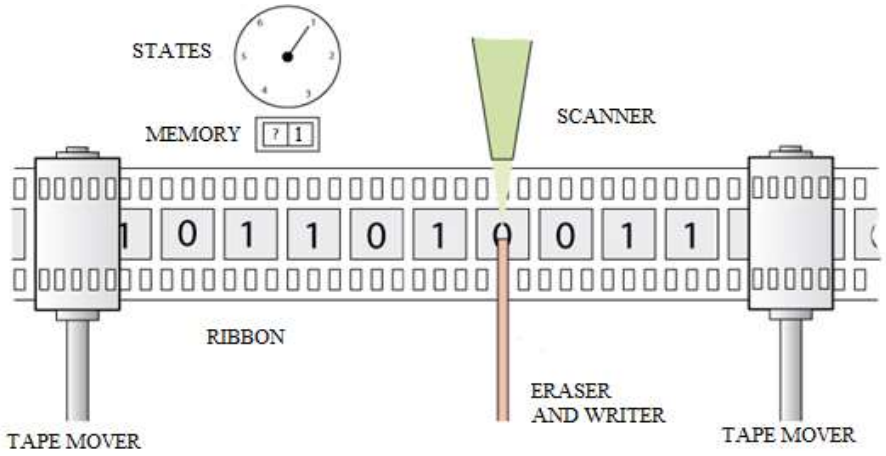
Un système de codage binaire



Les lettres et les nombres	Les opérations et états
A → 1 → 001	Droite → D → 4 → 100
B → 2 → 010	Écrire → E → 5 → 101
C → 3 → 011	Effacer → F → 6 → 110
D → 4 → 100	Gauche → G → 7 → 111
E → 5 → 101	Halte → H → 8 → 1000
…	…

Un système de simulation

Machine « A »

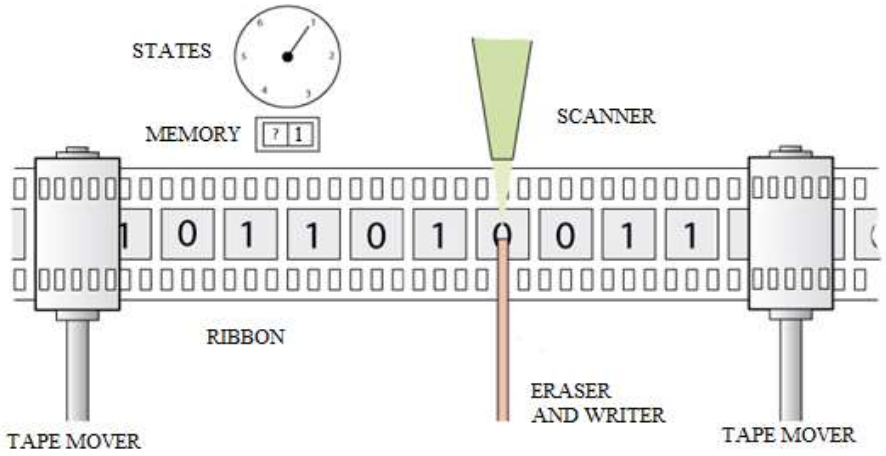


→

1	0	0	1	0	1	1	. . .

Description
Number (D.N)

Machine
Universelle



→

D.N(A)	0	1	. . .

La machine de Turing et ses limites

- **Machine conceptuelle** qui lit et écrit des **symboles** en suivant des **règles prédéfinies**.

  → Action limitée à ses instructions = son **code**.

- Langage universel = le **binaire**.

  → **Boucle** si instructions **contradictoires** (*Entscheindungsproblem*).

La machine de Turing préfigure les théories de l'IA

- Le comportement d'un individu = **stimulus-stress** (Behaviorisme, 1938).
- Stimulus-stress de l'organisme sont traduisibles par **0** et **1** (Cybernétique, 1947).
- L'activité neuronale = synapses **inhibées-activées** (Cognitivisme, 1956).

Paradigme du « **all-or-nothing** » de **McCulloch & Pitts**, *Logical Calculus of the Ideas Immanent in Nervous System*, 1943.

De la machine de Turing à l'IA

1. L'être humain est limité dans le temps, donc ses actions et pensées existent en un nombre **fini**.

Donc… ?

De la machine de Turing à l'IA

1. L'être humain est limité dans le temps, donc ses actions et pensées existent en un nombre **fini**.
2. Or, une machine de Turing universelle peut **simuler** toutes machines dont le comportement est réductible à un nombre **calculable**.
3. Donc, l'être humain peut être **représenté** par une machine de Turing.

Computing Machinery and Intelligence, 1950

- Théorie de la **machine-enfant**.
- Fait écho à **J. Piaget**, *La représentation du monde chez l'enfant* (1928).
- Évolution génétique = **apprentissage non-supervisé**.

Le jeu de l'imitation et le test de Turing

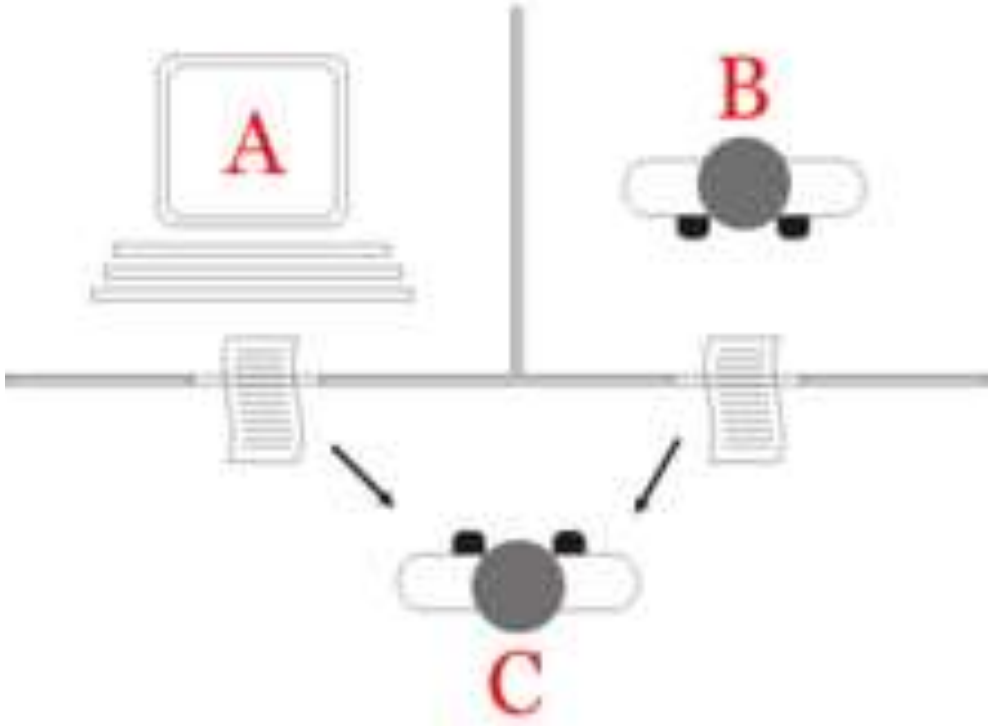
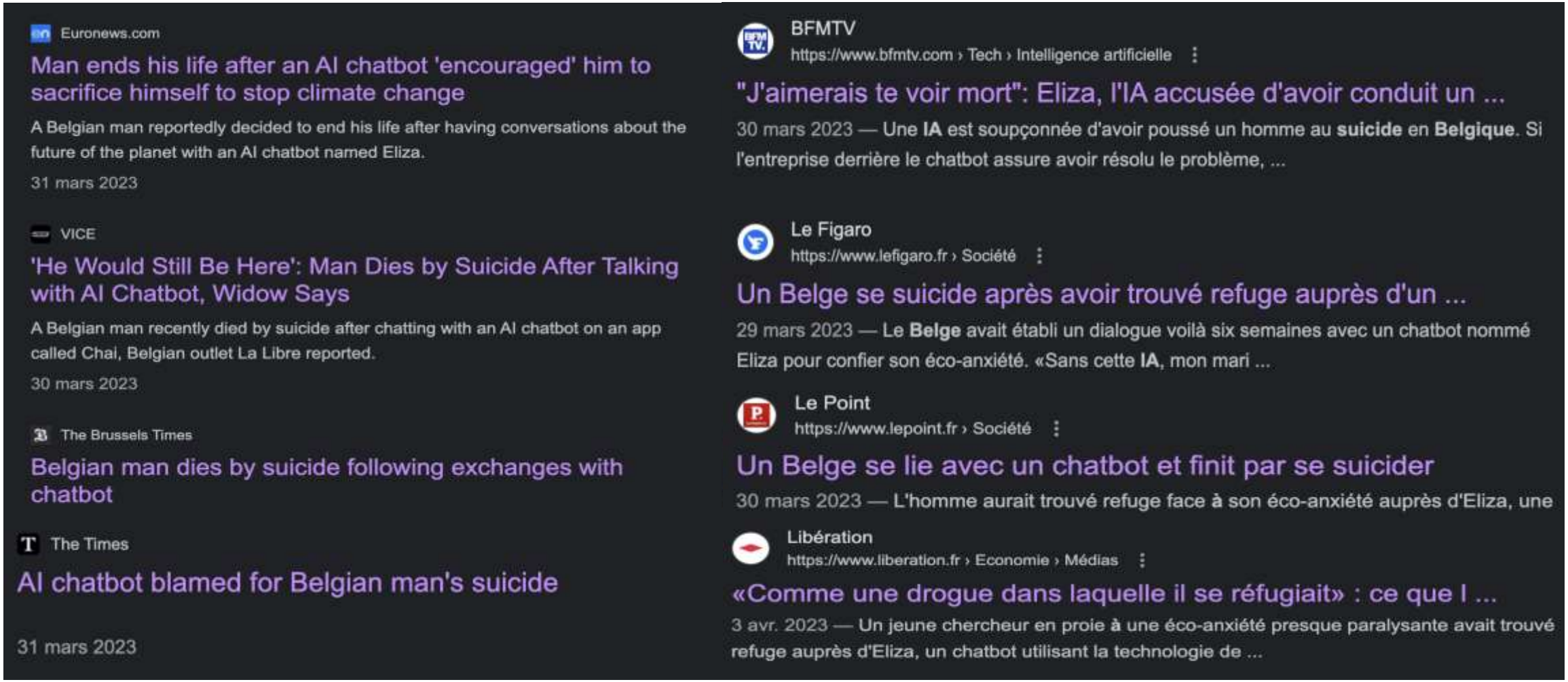


Figure 2. Situation du jeu de l'imitation
(A. P. Saygin, I. Cicekli, V. Akman, *Turing Test: 50 Years Later*, Minds and Machines, 10 (4): 463–518, 2000

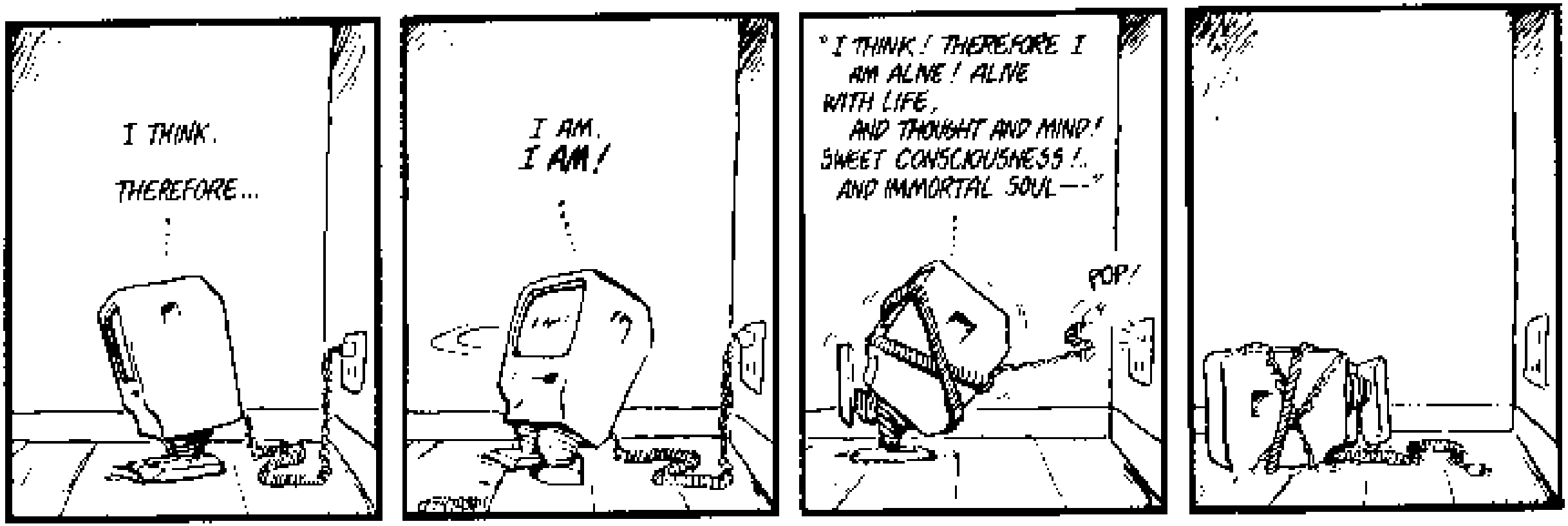
Une machine peut-elle penser ?

Source image: https://www.courrierinternational.com/article/intelligence-artificielle-google-renvoie-l-ingenieur-qui-disait-que-son-ia-etait-douee-de-conscience, 23 juillet 2022

Ne pas laisser la machine penser à sa place

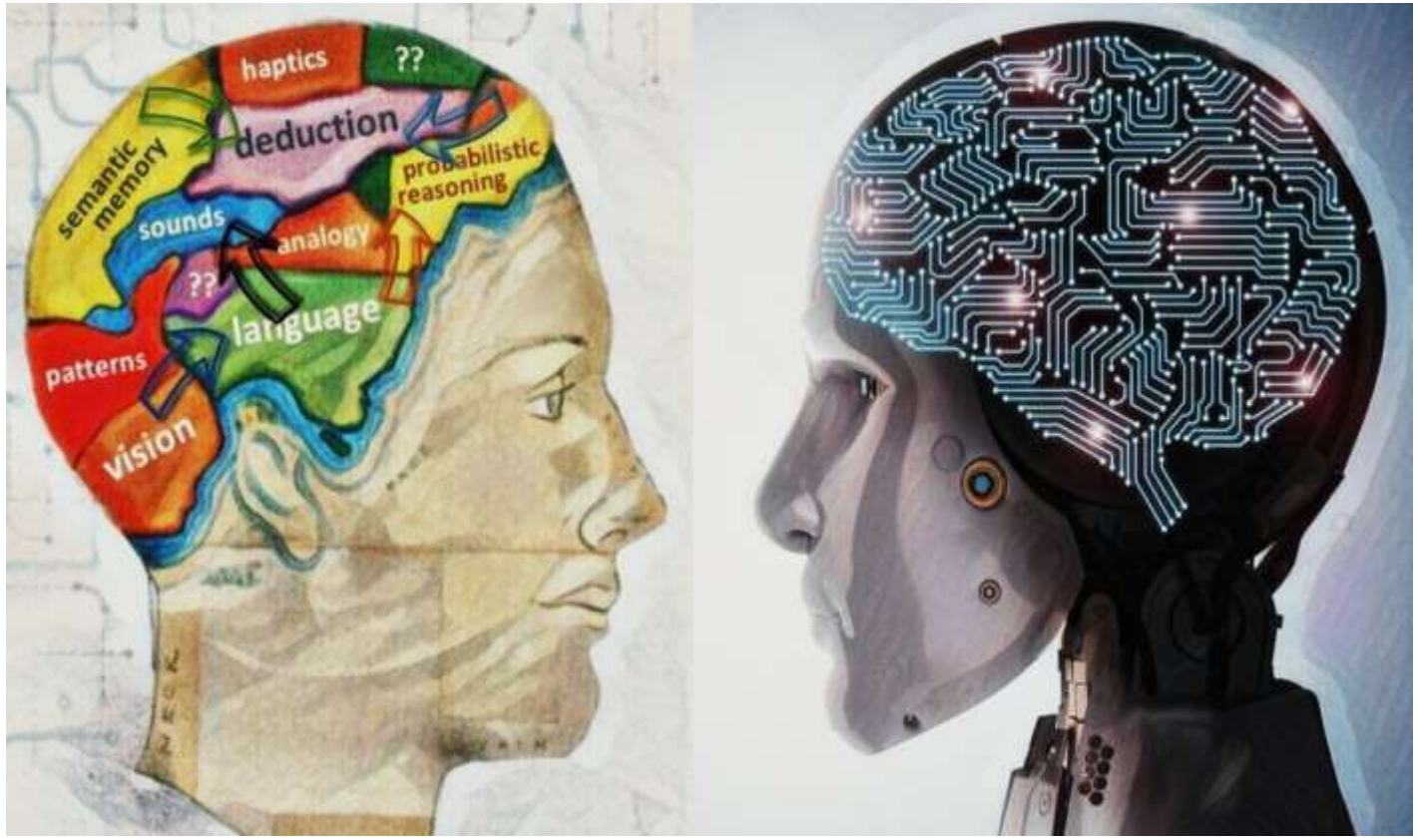


II. Débats philosophiques autour de l'IA



Tournant cognitiviste de 1950-1980

- **Conférence de Dartmouth** en 1956.
- Naissance du terme « **Intelligence Artificielle** » et des **sciences cognitives**.
- Hypothèse de la **modularité de l'esprit** par Fodor, 1975.
- L'IA sert désormais de modèle pour expliquer l'humain.



Source : https://arcingmind.com/2023/01/17/the-modular-mind/

Tournant cognitiviste de 1950-1980

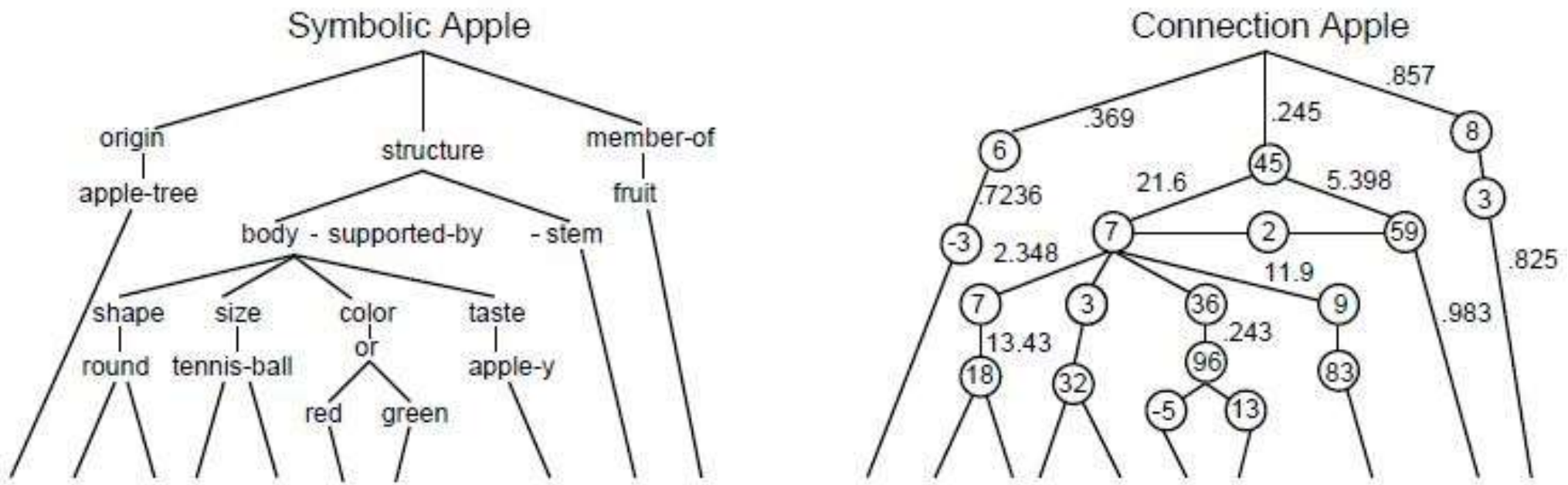


Figure 3. Deux modèles d'IA : symbolique et connexionniste.
(M. Minsky, *Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy*, AI Magazine Volume 12 Number 2, 1991)

1ère Critique : l'être-au-monde de la machine

Hubert Dreyfus, 1929 - 2017

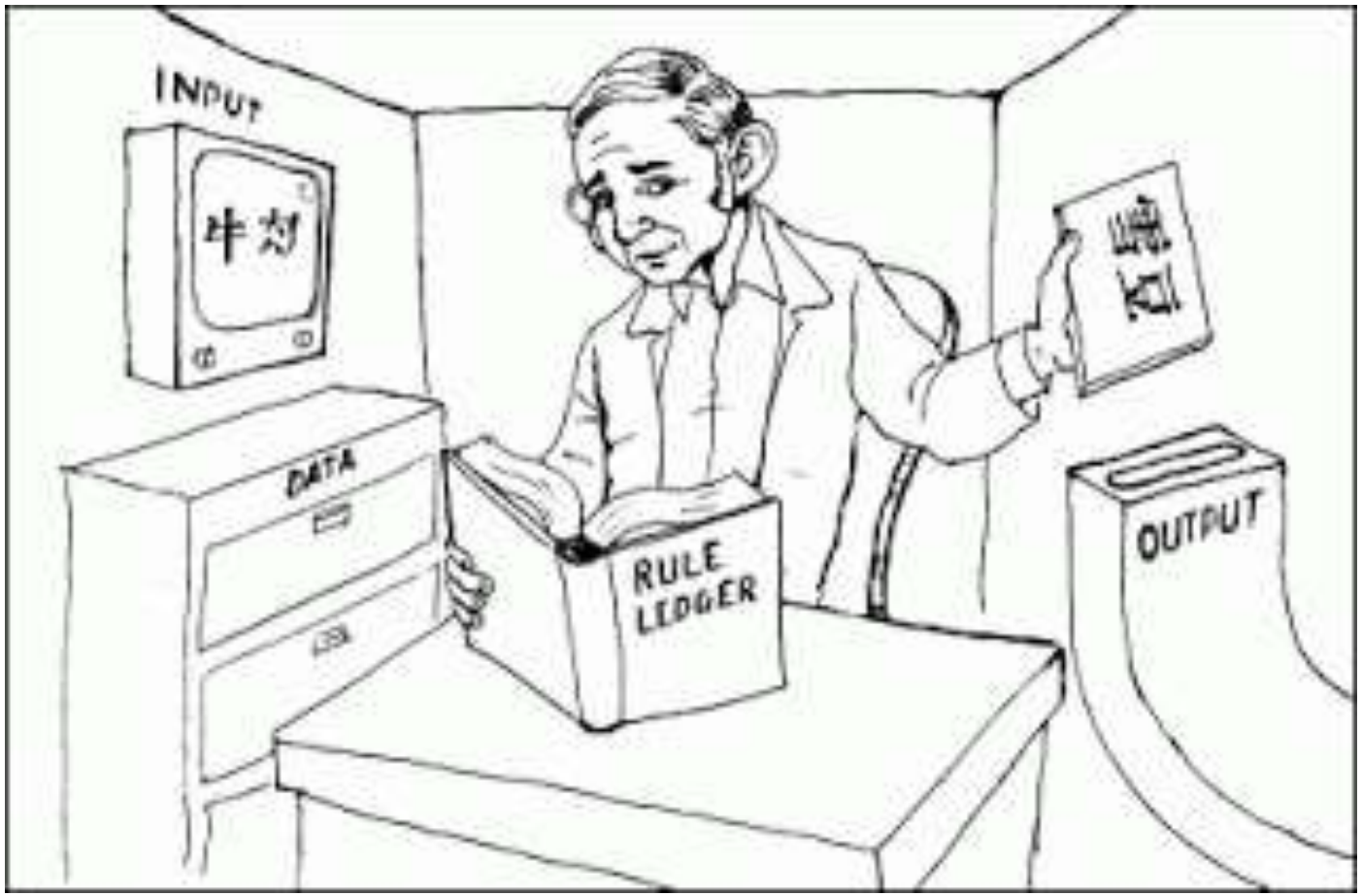
- *Alchemy and Artificial Intelligence* (1964)
- *What Computers Can't do: The limits of Artificial Intelligence* (1972)
- Introduit la **phénoménologie** européenne en Amérique du Nord pour critiquer l'IA.

1ère Critique : l'être-au-monde de la machine

Hubert Dreyfus, 1929 - 2017

- IA est fondée sur des présupposés **biologiques** et **psychologiques**.
- **Différence ontologique :** la machine n'est jamais un être « situé dans le monde ».
- Rien ne fait **sens** pour la machine.
- Réalité pas **formalisable**.

2e Critique : le langage de la machine



John Searle, 1932 -

- La « Chambre chinoise », *Minds, Brains & Programs* (1980)

- Syntaxe =/= Sémantique
- Fonction =/= Intention
- Cohérence =/= Compréhension

- **IA faible** =/= **IA Forte**

3e Critique : la guerre des intelligences

Bernard Stiegler, 1952 - 2020

- *De la misère symbolique* (2004).
- **Malaise social** dû à l'accélération technique.
- **Dépropriation** de nos fonctions cognitives.
- **Grammatisation** de nos relations sociales.
- Faits sociaux = Actes enregistrés.
- Disparition du processus d'**individuation** : Je = On

Complémentarité humain-machine

- **Paradoxe de Moravec (1980) :** ce qui est facile pour les machines est difficile pour l'humain et ce qui est facile pour l'humain est difficile pour les machines.

	Calculer de très grands nombres	**Comprendre le sens d'une phrase**	**Interpréter le contexte**	**Traiter la négation**	**Justifier une action**	**Lire des signes**
Humain	Difficile	Facile	Facile	Difficile	Difficile	Facile
Machine	Facile	Difficile	Difficile	Difficile	Facile	Facile

Y a-t-il un fantôme dans la machine ?

- La machine est faite à **notre image**.
- Mais pas d'**esprit** dans le circuit imprimé.
- Intelligence artificielle = Reflet de notre **intelligence sociale**.
  - Exemple du Chatbot « Tay » de Microsoft, 2014.

Image générée par IA.

Être exemplaire pour l'IA

« Ce qui réside dans les machines, c'est de la **réalité humaine**, du **geste humain** fixé et cristallisé en structures qui fonctionnent. Ces structures ont besoin d'être **soutenues** au cours de leur fonctionnement, et **la plus grande perfection** coïncide avec la plus grande ouverture, avec **la plus grande liberté de fonctionnement**. »

- **Gilbert Simondon**, *Du mode d'existence des objets techniques*, Paris, Éditions Aubier, 2012, (1958), p. 12

Être exemplaire pour l'IA

- Être « **à l'écoute** » des machines = comprendre les **individus** et **milieux techniques**.
- L'**éthique de l'IA** concerne davantage les humains que les machines.
- **Technophobie** et **technophilie** se nourrissent de l'ignorance de la **réalité de l'objet technique**.

Bibliographie

Anders G., *L'Obsolescence de l'homme : Sur l'âme à l'époque de la deuxième révolution industrielle*, (1956), Éditions de l'Encyclopédie des Nuisance, 2002

Dreyfus H., *What Computer Can't Do: A Critique of Artificial Reason*, Harper & Row Publisher, New York, 1972

Fodor J., *The Language of Thought*, Series Editor, Harvard University Press, 1975

Girard J.-Y., **Gödel K.**, **Nagel E.**, *Le théorème de Gödel*, Éditions du Seuil, 1989

Girard J.-Y., **A. M. Turing**, *La machine de Turing*, Éditions du Seuil, 1995

Leibnitz G. W., *Explication de l'arithmétique binaire, qui se sert des seuls caractères O et I avec des remarques sur son utilité et sur ce qu'elle donne le sens des anciennes figures chinoises de Fohy*, Mémoires de mathématique et de physique de l'Académie royale des sciences, Académie royale des sciences, 1703. ffads-00104781

Piaget J., *La représentation du monde chez l'enfant*, (1947), PUF, coll. Quadrige, 2013

Pinker S., *L'Instinct du langage*, (1994), Odile Jacob, 2013

Simondon G., *Du mode d'existence des objets techniques*, (1958), Éditions Aubier, 2012

Stiegler B., *De la misère symbolique, Vol. 1 : L'époque hyperindustrielle*, (2004), Éditions Galilée, Champs essais, Flammarion, 2013

Wiener N., *La Cybernétique : information et régulation dans le vivant et la machine*, (1948), Éditions du Seuil, 2014

- *God & Golem Inc. : Sur quelques points de collision entre la cybernétique et la religion*, (1964), Éditions de l'éclat, 2000

Bibliographie

McCarthy J., **Minsky M. L.**, **Rochester N.**, **Shannon C. E.**, *A Proposal For The Dartmouth Summer Research Project on Artificial Intelligence*, August 31 1955, http://www formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

McCarthy J., **Patrick J. Hayes P. J.**, *Some Philosophical Problems from the Standpoint of Artificial Intelligence*, Computer Science Department, Stanford University, 1969

McCulloch W. S., **Pitts W.**, *A Logical Calculus of the Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics, Vol. 5, pp. 115-133 (1943), reprinted in Bulleting of Mathematical Biology, Vol. 52, No. ½, pp. 99 -115, Society for Mathematical Biology, 1990

Minsky M., *Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy*, AI Magazine Volume 12 Number 2, 1991

Newell A., **Simon H.**, *Computer science as empirical inquiry: symbols and search*, Communication of ACM.19. 113-126, 1976

Saygin A. P., **Cicekli I., Akman V.**, *Turing Test: 50 Years Later*, Minds and Machines, 10 (4): 463–518, 2000

Searle J. R., *Minds, Brains and Programs*, The Behavorial and Brain Sciences, Vol. 3, Cambridge University Press, 1980

Turing A. M., *On Computable Numbers, with an Application to the Entscheidungsproblem*, Proceedings of the London Mathematical Society, London Mathematical Society, 1936

- *Intelligent Machinery*, National Physical Laboratory, 1948
- *Computing Machinery and Intelligence*, Mind, 59, 433-460, 1950

Weizenbaum J., *ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine*, Massachusetts Institute of Technology, Cambridge, Mass., Vol. 9, No. 1, January, 1966

Merci pour votre attention