La IA desde la filosofía de la mente

Antecedentes, desafíos, críticas

RODRIGO ALFONSO
GONZÁLEZ FERNÁNDEZ
rodgonfer@gmail.com

CENTRO DE ESTUDIOS COGNITIVOS
UNIVERSIDAD DE CHILE
8 DE AGOSTO DE 2024

Alan Turing

Creo que al final del siglo el uso de las palabras y la opinión de la gente educada habrán cambiado tanto que uno será capaz de hablar de máquinas pensantes sin esperar incurrir en contradicción…

Marvin Minsky

La IA puede definirse como una disciplina encargada de crear máquinas programadas que sean capaces de hacer cosas que requieren la misma inteligencia que si fuesen hechas por los humanos…

INTRODUCCIÓN

Motivación: es común hablar de la IA, pero muchas veces se lo hace desde un punto de vista técnico. ¿Puede la filosofía decir algo? Esto es, ¿hay una filosofía de la IA? ¿Por qué la reflexión filosófica es relevante en relación con la IA?

Hay problemas en la IA que son filosóficos, en particular, uno crucial:

¿Es la mente un computador? Y, a la inversa: ¿son los computadores mentes?

Por otra parte, hay temas de la IA que se cruzan con disciplinas filosóficas contemporáneas:

1. La filosofía de la mente contemporánea
2. La ética de la IA

OUTLINE

1. Los albores de la IA: Babbage y la sombra de Descartes

2. Turing y sus dos aportes a la IA: la Máquina de Turing y TT

3. Searle y la pugna teórica con la IA fuerte (TCM en Block)

4. Las nuevas tendencias en la IA: Sobre cómo superar el dictum y su supuesto metafísico (¿Solo una disputa académica?)

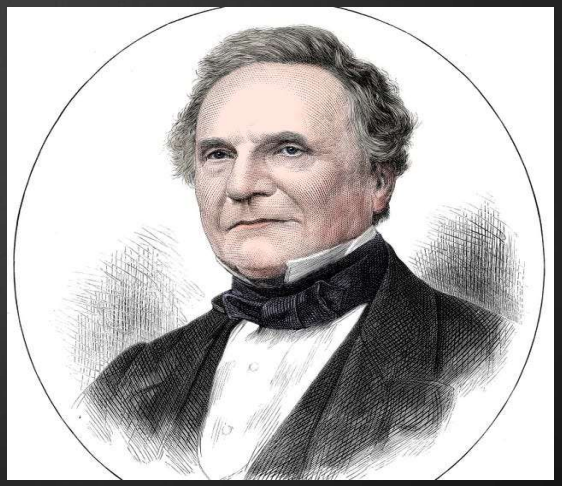
5. Algunos desafíos éticos de la IA futura

1. Los albores de la IA: Babbage y la sombra de Descartes

- Hay dos figuras prominentes en el albor de la IA, en relación con su imposibilidad y posibilidad en principio, respectivamente

- Descartes y su “dictum”

Charles Babbage y sus máquinas



(…) Si hubiese máquinas tales que tuviesen los órganos y figura exterior de un mono o de otro cualquiera animal, desprovisto de razón, no habría medio alguno que nos permitiera conocer que no son en todo de igual naturaleza que esos animales; mientras que si las hubiera que semejasen a nuestros cuerpos e imitasen nuestras acciones, cuanto fuere moralmente posible, *siempre tendríamos dos medios muy ciertos para reconocer que no por eso son hombres verdaderos*;

El primero es que nunca podrían hacer uso de palabras ni otros signos, componiéndolos, como hacemos nosotros, para declarar nuestros pensamientos. [Una máquina] no se concibe que ordene en varios modos las palabras para contestar al sentido de todo lo que en su presencia se diga, como pueden hacerlo aun los más estúpidos de entre los hombres

El segundo es que, aun cuando hicieran varias cosas tan bien y acaso mejor que ninguno de nosotros, no dejarían de fallar en otras, por donde se descubriría que no obran por conocimiento, sino sólo por la disposición de sus órganos, pues mientras que la razón es un instrumento universal, que puede servir en todas las coyunturas, esos órganos necesitan una particular disposición para cada acción particular.

René Descartes, Discurso del Método, 5ª parte, pp. 33-34

¿Qué supuesto metafísico subyace al dictum?

Dualismo cartesiano

1. La mente es *separable* del cuerpo (*no está separada*, sino que puede separarse. ¿Cómo? Mediante Dios y lo concebible en el entendimiento)

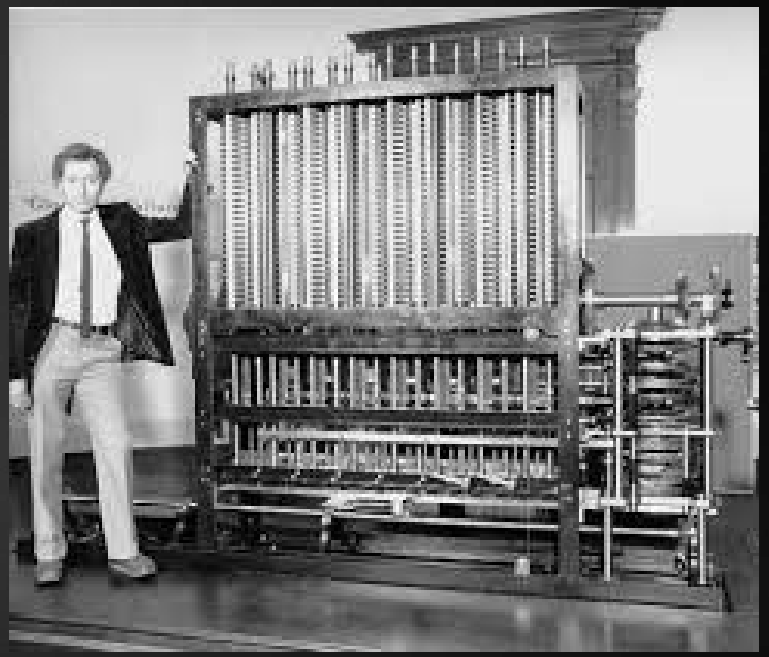
2. Una máquina es *una cosa física* metafísicamente incompatible con la mente. Mientras que esta es inmaterial, indivisible e inmortal, la primera es material, divisible y corruptible

Paradójicamente, el *funcionalismo filosófico* es compatible con el dualismo cartesiano, porque supone que hardware y software son separables. Por ejemplo, la función del computador es diferente de la realización material de este → Principio de realizabilidad múltiple

Esto trae consecuencias para las aproximaciones alternativas de la ciencia cognitiva (que veremos en la sección 4)

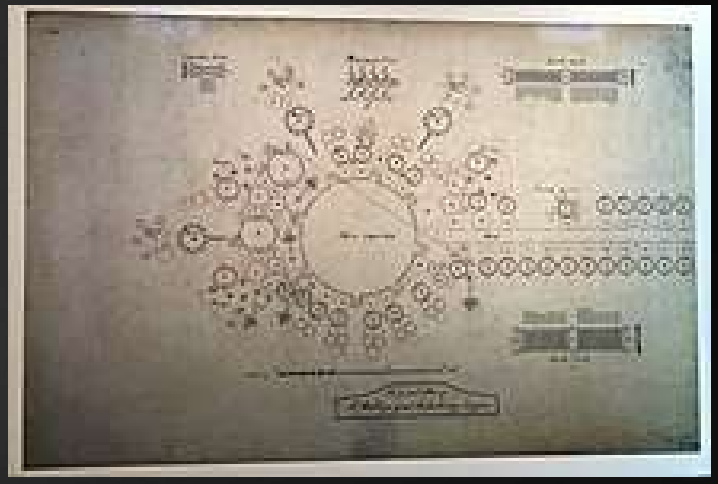
¿Cómo se inició la IA? El problema de las tablas de cálculo…

1. Siglo XIX era del maquinismo e industrialización
2. Las tablas de cálculo y los "computadores" del siglo XIX
3. Un problema práctico para evitar pérdidas humanas y económicas
4. La solución de Babbage: Máquina de las Diferencias
5. Problemas para construir las máquinas: costos, problemas técnicos, problemas de patentes (Caso de Joseph Clement)
6. Las mejoras técnicas de la máquina, segunda versión
7. El motor analítico y el genial descubrimiento de Augusta Ada Condesa de Lovelace



En relación con los adelantos de las máquinas de Babbage y el uso de vocabulario psicológico para describer su funcionamiento…

Para oponerse al dictum cartesiano, Babbage propuso que la inteligencia de máquina era posible. En particular, sostuvo que el uso de vocabulario psicológico era necesario para evitar largos circunloquios al describir el funcionamiento de la máquina de las diferencias y sus sucesoras. En concreto, planteó que



[…] Con el principio de los mecanismos de reserva sucesivos, se me ocurrió que podría ser que se le enseñase a los mecanismos a *realizar* procesos mentales, a saber, anticipar … La idea se me ocurrió en octubre de 1834. Me significó pensar mecho, pero logré llegar al principio subyacente en corto tiempo. Luego de llegar a él, el paso siguiente fue enseñarle al mecanismo que anticipaba a actuar en función de dicha anticipación. Esto no fue tan difícil; se construyeron algunos mecanismos que, aunque no eran para nada simples, hicieron posible pensar que la máquina podia ser construída (Babbage 2010, pp. 104-105).

Para Babbage: La inteligencia de máquina *está* en la máquina…

2. Turing y sus dos aportes a la IA: la Máquina de Turing y TT

2.1 La Máquina de Turing

- Máquina de Turing (TM): a diferencia de Babbage, quién piensa que la inteligencia está *en* la máquina, Turing (1936) propone que la inteligencia es producto de un algoritmo (la máquina solo ejecuta el algoritmo; de hecho, el algoritmo es un programa, una máquina con base en la recursión)

- TM es fundamental para definir qué es computar: es la aplicación de un algoritmo a un problema. Mediante éste y los pasos finitos del algoritmo se llega a un resultado

- Algoritmos y TMs son conceptos interrelacionados: un algoritmo es un procedimiento que puede implementarse mediante TM y TMs no hacen sino implementar algoritmos para alcanzar resultados

- Pero, ¿qué es un algoritmo? Es un conjunto de reglas específicas que, aplicadas a un problema, generan una solución mediante pasos finitos y recursión

- Los algoritmos fueron popularizados por el matemático persa Abu Ja'Far Mohammed ibn Mûsâ al-*Khowâzarim* cerca del 825 DC. Pero existían desde mucho antes…

- Por ejemplo, el algoritmo de Euclides: encontrar el máximo común denominador de dos números enteros.

- Algoritmo con reglas y pasos finitos, siendo paso iii recursivo:

    i) Dividir dividendo y divisor, anotando cociente y resto (R);
    ii)Si R = 0, *halt*;
    iii)Si R ≠ 0, tomar divisor y resto anteriores para ejecutar paso 1.

- Por ejemplo, 99 y 15.

Dividendo	Divisor	Cociente	Resto
99	15	6	9
15	9	1	6
9	6	1	3
6	3	2	0

- Luego, el MCD entre 99 y 15 es 3

• Lo que uno hace mediante estos cálculos también lo puede hacer una TM, porque esta resuelve problemas aplicando las reglas del algoritmo mediante *descomposición recursiva* y llegará al mismo resultado. ¿Qué es la descomposición recursiva?

Por ejemplo, puedo descomponer la multiplicación recursivamente en sumas y restas así:
M x N = A

Reglas:
i) Sumar 1 vez M a A y restar 1 a N;
ii) Si N = 0, halt;
iii) Si N ≠ 0, ejecutar paso i

- Por ejemplo, 3 x 3 →

M	x	N	=	A
3		2		3
3		1		6
3		0		9

- Tal programa opera automáticamente, recursivamente, y mediante la descomposición recursiva de un problema complejo: la multiplicación…

- ¿Y qué pasa con una TM? Todas sus instrucciones se encuentran expresadas en una tabla de máquina. En este caso, las acciones están descritas por las reglas del siguiente programa (en quíntuplos: dos operaciones de input, tres de output):

- Una TM que sume 2 + 3 = 5

	q_0	q_1
1	1D q_0	**# Halt**
+	1D q_0	
#	**# I q_1**	

#	#	1	1	1	+	1	1	#	#	
		↑q_0								

- Si las instrucciones de TM se encuentren en la cinta, esta TM es una UTM (máquina universal de Turing). Dichas instrucciones son el input inicial de la máquina. Tal UTM puede imitar perfectamente el comportamiento de *cualquier* otra TM. ¿Es la mente una UTM? Turing cree que sí:

- Entrevista radial a Turing para BBC en 1951:

> Para lograr que nuestro computador imite a una máquina sólo es necesario programarlo para que calcule lo que la máquina en cuestión haría bajo ciertas circunstancias […] Ahora bien, si una máquina en particular puede describirse como un cerebro, tenemos que solamente programar nuestro computador digital para imitarlo y también será un cerebro. Si se acepta que los cerebros reales, descubiertos en animales, y en especial en el hombre, son una clase de máquina, se seguirá entonces que *nuestro computador digital, debidamente programado, se comportará como un cerebro.* Este argumento presupone una idea que puede ser razonablemente cuestionada […] que esta máquina debiera ser de una naturaleza cuya conducta sea en principio predecible mediante cálculo […] Nuestro problema es, entonces, cómo programar una máquina para imitar al cerebro, o si lo pudiésemos expresar de una manera más breve y menos rigurosa, *para que piense* (En Copeland 2003, p. 11, énfasis mío).

2.2 TT: Un método empírico para probar que una máquina programada es una mente

¿Pueden pensar las máquinas? (La pregunta del dictum cartesiano)

Turing: evitemos el problema conceptual mediante el reemplazo de la pregunta (el dictum)…

¿Por qué? Una discusión de conceptos lleva al uso de ellos, y a una encuesta tipo Gallup. Luego, la pregunta se puede reemplazar por el juego de la imitación.

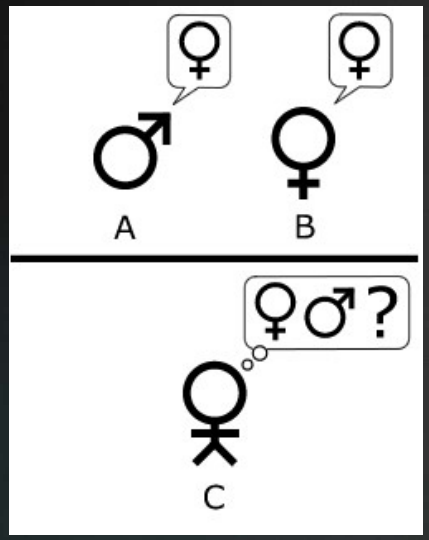
¿Existe algún problema con el funcionamiento de la máquina programada?

Sí, Turing anticipa una objeción sólida así:

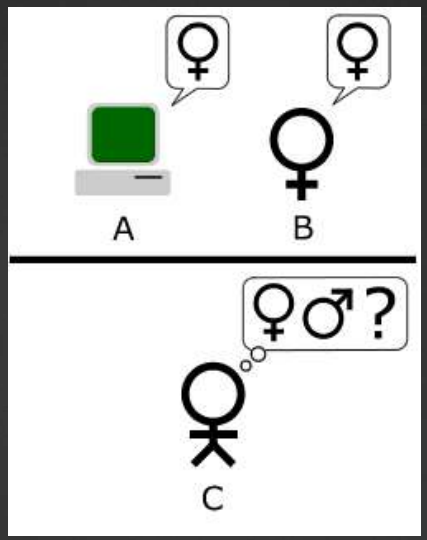
¿Podría ser que las máquinas hagan algo que debería ser descrito como pensamiento, pero que es diferente de lo que hace el ser humano? Esta objeción es sólida, pero a menos podemos decir que si, con todo, una máquina puede construirse para desempeñarse bien en el juego de la imitación, no debemos preocuparnos de la objeción (Turing 1950, p. 42, volveremos a esta objeción más adelante).

Las tres versiones del juego de la imitación en TT: la esencia del funcionalismo de Turing

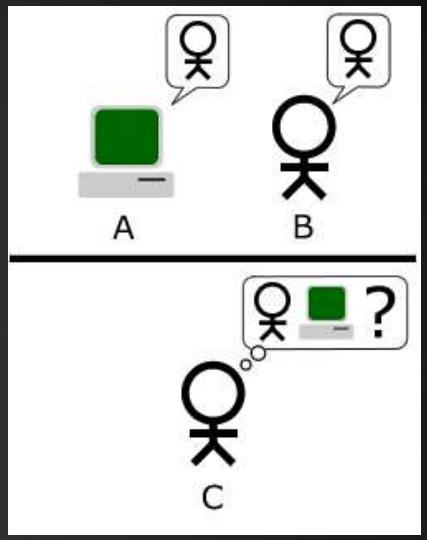
Versión 1



Versión 2



Versión estándar



Preguntas y respuestas de jueces humanos…interrogatorios de 5 minutos

Hacerse pasar por X, el reemplazo de X mediante conducta lingüística, como el elemento central del test, y *central de la IA*. Imitar a X no requiere de propiedades físicas de X

Es decir, Turing adscribe al principio de realizabilidad mútiple del funcionalismo

La predicción de Turing…

"Creo que en 50 años será posible programar computadores con una capacidad de 10^9 para hacerlos jugar el juego de la imitación tan bien que el interrogador promedio no tendrá más de 70% de chance de hacer la identificación correcta después de 5 minutos de preguntas. La pregunta original "¿puede pensar una máquina?" es tan sin sentido que no merece discusión" (Turing 1950, p. 49)

Turing y Descartes, la obsesión con el lenguaje como evidencia de la existencia de mente → Chatbots: ELIZA, PARRY, SHDRLU, etc.

Eugene Goostman pasó el test en 2014, Chat GPT 4 lo hizo en 2024

¿Por qué no estamos todavía convencidos del todo de que la mente es solo un computador? Por la siguiente pregunta (filosófica): ¿*Entiende* Chat GPT4 lo que responde?

TT: ¿Es lo epistémico, verificar conducta linguistica inteligente, suficiente para tener mente, una cuestión metafísica, tal como Turing cree?

Resultado de las tesis de Turing: TM y TT

+

Otros antecedentes de construcción de computadores (Z3, ENIAC, BINAC, Mark I, etc.) y de creación de programas como el Lógico Teórico (LT)

¡Nace la IA como disciplina en 1956! Hubo dos hitos cruciales para su nacimiento…

1. LT probó algunos Teoremas de Principia Mathemática de Russell y Whitehead:

Dado X o Y = V, y que Y = F, entonces X = V

2. Dartmouth Summer Research Project on Artificial Intelligence: organizado por John McCarthy, conferencia no exitosa, pero dio sentido de identidad a la disciplina

3. Searle y la pugna teórica con la IA fuerte (TCM en Block 1991)

- ¿Son los programas de la IA mentes, o son solo instrumentos para lograr entender la cognición humana?
- Teniendo presente programas como SAM (Script Applier Mechanism) de Schank y Abelson, Searle (1980) hace una distinción entre:

1. IA *fuerte*: es la tesis de que los programas computacionales son mentes, y que los programas explican la existencia de estados mentales → Turing y otros

2. IA *débil*: los programas de la IA son meros instrumentos que nos permiten explicar la cognición humana, son *como-si* mentes

Para mostrar que IA fuerte es una teoría falsa, a la que subyace una tesis errónea, Searle propone un experimento mental, la Habitación China, que engaña al 100% de los interrogadores tipo TT

En concreto, en este EM Searle se propone una operación para responder esta pregunta: ¿Qué pasaría si mi mente operase de acuerdo con los principios de la IA fuerte?

La pregunta previa intenta responder esta otra: ¿Es la sintaxis (las propiedades sintácticas de un programa) *suficiente* para crear una mente?

En particular, ¿es la manipulación de símbolos con base en reglas sintácticas suficiente para crear (causar) una mente?

Tres buenas objeciones a la Habitación China

i) La réplica del sistema: Searle más el libro de reglas, los bancos de datos y la manipulación simbólica entienden chino. La replica de Searle: si la persona memorizase todos los elementos de la habitación, no entendería chino.

ii) La réplica del robot: si un programa operase un robot, permitiendo que interactuase simbólicamente con el ambiente (para que existieran estados intencionales), dicho robot actuaría con el ambiente a través del input recibido y los dispositivos perceptuales. ¿Entendería ese robot? Sí: Fodor 1980, Dennett 1980.

iii) La réplica de los Churchlands (1990): Searle no puede saber si el axioma 3 de su argumento, que la sintaxis ni es constitutiva ni sufiente para una mente, es verdadero. Además, el conexionismo es una alternativa plausible…

¿Por qué son buenas objeciones? Porque cambian el foco del análisis: la realización física de los programas, conectados con el ambiente, podría permitir la existencia de mente, a diferencia de lo que propone Descartes con su dictum dualista. Esto da origen a un programa de investigación alternativo, lo que llamo las teorías “alternativas” de la cognición.

Pero antes, ¿Hay una vía de escape de la Habitación China?

Parece que sí, según Turing…Una vía pragmática…

Cuando habla sobre la objeción de la conciencia al Juego de la Imitación, afirma lo siguiente:

"De acuerdo con la objeción de la conciencia, la única mánera de saber si alguien piensa es ser como esa persona. Este es el punto de vista del solipsista. Es una visión lógica, que hace difícil la comunicación. A puede llegar a creer "A piensa pero B no" mientras que B cree "B piensa mientras que A no". En vez de polemizar acerca de este punto, es más usual adoptar la educada convención de que todo el mundo piensa" (Turing 1950, p. 52).

Computadores y mentes: ¿Una discusión solo filosófica *(académica)*?

4. Las nuevas tendencias en la IA: Sobre cómo superar el dictum y sus supuestos metafísicos

Primero, el dictum cartesiano se apoya en una teoría metafísica sobre la mente, el dualismo

Segundo, el funcionalismo, compatible con el dualismo, se apoya en el principio de realizabilidad múltiple

Tercero, las réplicas (seleccionadas) cuestionan la separabilidad mente/cuerpo

¿Cuál es el resultado?

I) Mente corporizada (cognición corporizada)

A) Los procesos mentales no son computacionales

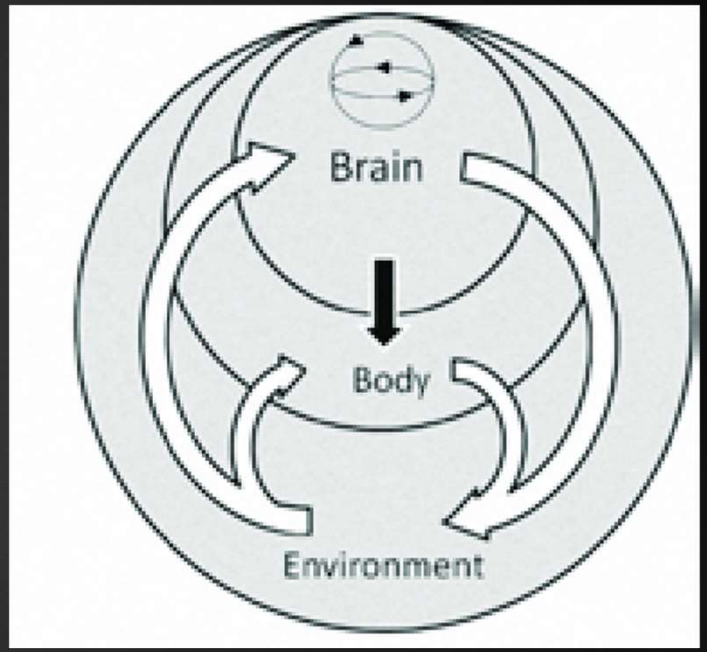
B) El cerebro no es el fundamento principal de la cognición

C) Lo crucial: la corporalidad del agente cognoscente causa las habilidades cognitivas

En síntesis, el cuerpo o la interacción corporal constituyen y contribuyen a explicar la cognición, tesis que requiere de un nuevo paradigma de investigación en la ciencia cognitiva

- **II) Mente incrustada [*embedded*]**
- A) La mente incrustada sigue las mismas directrices de la cognición situada (externista)
- B) La mente es dependiente de la interacción de cuerpo y ambiente
- C) Ello da lugar a que la enseñanza, el aprendizaje y la instrucción sean *situados*

- En síntesis, la mente incrustada considera, a diferencia del paradigma clásico, que mente, cuerpo y cognición/aprendizaje *no son separables*



III) Mente extendida (Clark y Chalmers)

A) La mente se extiende más allá de la caja craneana, ¿dónde?

B) En la lista de compras, en la memoria del PC, en el celular, etc.

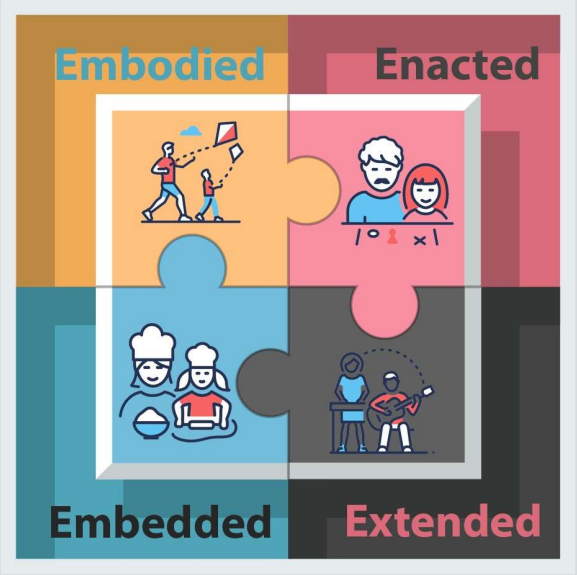
C) Hay muchos dispositivos en el ambiente que pueden ser parte integrante de la mente

En síntesis, el dualismo cartesiano afecta la demarcación de los problemas de la ciencia cognitiva (lo que aplica para I y II también). Por tal motivo, es conveniente rechazar el paradigma clásico de que la mente es solo un computador, porque dicha mente no resulta separable del ambiente

Las teorías alternativas se resumen en la *4E cognition*:

Cognition is grounded in our senses and concrete physical experience.

Cognition is for goal-directed action in the real world.



Cognition is woven into culture. Learning has a social context.

Cognitive systems include tools, devices and the people around us.

5. Algunos desafíos éticos de la IA futura

- De acuerdo con Luciano Floridi (2021), hay 5 principios éticos que deben regir la IA:
- i) Beneficencia para la humanidad: Bienestar de la humanidad, bien común, dignidad y sostenibilidad
- ii) No maleficencia: Evitar uso incorrecto o excesivo de la IA, e.g., prevención de atentados a la privacidad de las personas
- iii) Autonomía: Balancear entre toma de decisiones que humanidad debe conservar, y la que puede delegar a máquinas
- iv) Justicia: promoción de la prosperidad, la solidaridad, e.g., evitar injusticias como la discriminación
- v) Explicabilidad: sistemas de IA deben ser transparentes e inteligibles para poder adjudicar responsabilidad
- I a IV provienen de la bioética, V es una propuesta de Floridi

6. Referencias principales

- Block, Ned (1991), "The computer model of the mind." In: D.N. Osherson and E.E. Smith (eds.) *Thinking: An Invitation to Cognitive Science*, Vol. 3. Cambridge, Mass.: MIT Press, pp. 247-89.
- Churchland P.M. and Churchland P.S. (1990), "Could a machine think?" *Scientific American* January 1990, pp. 26-31.
- Clark, Andy y Chalmers, David (1998), "The extended mind", *Analysis* 58, 1, pp. 7-19.
- Copeland, J. (1993), *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.
- Descartes, R. (2004), *Discourse on Method (Ch. 5)*. In: S. Shieber (ed.) *The Turing Test*. Cambridge, Mass.: MIT Press.
- Floridi, Luciano (2021), A unified framework of five principles for AI in society. In: *Ethics, governance, and policies in Artificial Intelligence.* Cham: Springer, pp. 5-17.
- González, R. (2015), "¿Importa la determinación del sexo en el Test de Turing?", *Aurora*, Vol. 27, 40, pp. 277-295.
- González, R. (2020), "Género, imitación e inteligencia: Una crítica del enfoque funcionalista de Alan Turing". En: P. López- Silva y F. Osorio (eds.) *Filosofía de la Mente y psicología Enfoques interdisciplinarios*. Santiago de Chile. Ediciones Universidad Alberto Hurtado, pp. 99-122.
- Searle, J. (1980), "Minds, brains and programs", *Behavioral and Brain Sciences* 3, 417-457.
- Swade, D. (2000), *The Difference Engine: Charles Babbage and the Quest to build the First Computer.* London: Penguin.
- Turing, A.M. (1936): "On computable numbers, with an application to the Entscheidungsproblem." Proceedings of the London Mathematical Society, series 2, Vol. 42, 231-65 (with corrections in Vol. 43: 544-6). Reprinted in: M. Davis (ed.) The Undecidable. New York: Raven Press.
- Turing, A.M. (1950), "Computing intelligence and machinery." *Mind* LIX, no. 2236, (Oct. 1950), 433-60. Reprinted in: M.A. Boden (ed.) *The Philosophy of Artificial Intelligence*. Oxford: OUP, pp. 40-66. Alianza Editorial, diversas reimpresiones