¿Puede la Inteligencia artificial dar una “mente” a las máquinas?

Hugo Leonardo Rufiner

Universidad Nacional del Litoral – Universidad Nacional de Entre Ríos - CONICET

Resumen

A lo largo de la historia la mente ha sido concebida y definida en diferentes formas. En la actualidad el funcionalismo ha asociado la dualidad mente-cerebro más fuertemente con los conceptos "equivalentes" de software-hardware provenientes del ámbito de las ciencias de la computación. La enorme evolución del software a través de la *inteligencia artificial* ha permitido emular varios aspectos de la inteligencia humana. Esto que parece nuevo, ha sido en realidad la motivación de esta disciplina desde sus inicios. Un hito importante en esta evolución son las técnicas bio-inspiradas de aprendizaje profundo que han posibilitado mejorar el desempeño de los algoritmos en diversas tareas y la implementación de sistemas prácticos de inteligencia artificial *de propósito general*. Sin embargo, ¿puede esta evolución llegar a dotar de una mente real a nuestras computadoras? ¿Podría por ejemplo dotarlas de una consciencia? A partir de la interacción con mi colega Francisco Soler se han suscitado varios escenarios posibles para estas preguntas que trataré de analizar en esta charla.

Introducción

La pregunta del título es bastante sugestiva y seguramente genera muchas expectativas por lo que me deja un guante difícil de levantar. Muchos conceptos complejos, aún fuera de nuestro completo conocimiento, y provenientes de muy diferentes disciplinas están en juego para intentar abordar una respuesta (Física, Biología, Medicina, Psicología, Psiquiatría, Filosofía, Teología, etc.). Mi formación original proviene de la ingeniería biomédica, especializado luego en temas de procesamiento de señales bioinspirado e inteligencia computacional, por lo que voy a tratar de responder cuando pueda desde ese lugar, esperando no caer en ningún ejercicio ilegal de la filosofía (o de cualquier otra disciplina diferente a la propia). Por supuesto que para comenzar lo primero que uno pensaría es en introducir algunas definiciones importantes desde las neurociencias, como por ejemplo que se entiende por mente, consciencia, inteligencia, etc. Sin embargo, en muchos casos no existe consenso dentro de la comunidad científica acerca de estas definiciones. Por ejemplo, la consciencia[1] no puede

[1] Los términos *conciencia* y *consciencia* no son intercambiables en todos los contextos. En sentido moral, como ‘capacidad de distinguir entre el bien y el mal’, solo se usa la forma *conciencia* por eso aquí preferimos usar la forma *consciencia* que es menos común y también menos ambigua para lo que nos ocupa en este trabajo.

definirse operativamente de una manera fácil, ya que es un fenómeno subjetivo y, por lo tanto, no se puede observar directamente. Se asocia este fenómeno complejo con la experiencia en primera persona, la interioridad, la subjetividad, la autopercepción, etc. De acuerdo con mi experiencia reciente como editor asociado en una revista del área, a pesar de que se han publicado más de 20.000 artículos científicos relacionados con la consciencia, todavía hay colegas que consideran este concepto como “no científico”, “motivado religiosamente” o “idea equivocada de la realidad”. Así que me remitiré sólo a una breve introducción de los términos principales sin entrar en detalles ni discusión. Además, esta charla se da en el marco de una discusión más amplia donde ya se han tocado muchos de estos temas y se tiene una serie de lecturas sugeridas como material de consulta adicional.

Se puede definir a la “mente” como el fenómeno responsable del entendimiento, el raciocinio, la percepción, la emoción, la memoria, la imaginación, la voluntad, la consciencia y otras habilidades cognitivas, muchas de las cuales son características esenciales del ser humano. El concepto de mente ha sido concebido en diferentes formas o categorías a lo largo de la historia: como una sustancia distinta del cuerpo, una parte, un proceso, o una propiedad. Las concepciones materialistas dominantes actuales engloban este concepto en la teoría de la identidad mente-cerebro y el funcionalismo. Es en este último sentido que la dualidad mente-cerebro se ha asociado más fuertemente con los conceptos "equivalentes" de software-hardware provenientes del ámbito de las ciencias de la computación. Sobre el uso de analogías de este tipo en la ciencia y en el contexto particular que nos ocupa, sus “peligros”, ventajas, alcances y limitaciones ya discutió mi colega (ver charla de Francisco Soler).

La Inteligencia Artificial (AI[2]) surge como disciplina hace más de medio siglo, cuando las Ciencias de la Computación eran todavía incipientes. Tampoco ha existido una definición única y uniforme para la Inteligencia Artificial a lo largo del tiempo. Esto se debe a que lo que consideramos como comportamientos inteligentes para una máquina también ha ido cambiando. Por ejemplo, en algún momento se suponía que si se creaba una máquina que pudiera realizar rápidamente cálculos matemáticos complejos eso podría considerarse como signo de inteligencia. John McCarthy definió la AI como: “La ciencia e ingeniería de hacer máquinas inteligentes” (1956). Más recientemente Bernstein y Curtis dan una definición más moderna: “El estudio y diseño de agentes inteligentes, donde un agente inteligente es un sistema que percibe su entorno y toma acciones que maximizan sus posibilidades de éxito” (2008).

Dentro de la AI surgieron también varios campos de investigación, tales como los Sistemas Expertos (ES), el Aprendizaje de Máquinas (ML) y las Redes Neuronales Artificiales (ANN). Estos últimos han cobrado mucha popularidad por su impacto en el desarrollo de tecnologías para nuestra vida cotidiana que hacen que nuestros dispositivos parezcan realmente inteligentes. Un hito decisivo en esta evolución es la reciente aparición de las técnicas de

[2] En lo que sigue utilizaremos las siglas en inglés debido a su difusión más amplia en la comunidad de Ciencias de la Computación.

Aprendizaje Profundo que han hecho posible la implementación, entre otras cosas, de sistemas prácticos de AI “General” (o de propósito general, GAI). Pero repasemos brevemente como llegamos a este punto y si es verdad que esta evolución puede culminar en la provisión de una mente real (con todas sus facultades) para nuestras computadoras. Esto es lo que parecen augurar varios profetas modernos de la AI que hablan de que estamos muy cerca de una “singularidad” tecnológica de consecuencias enormes.

¿Dónde estamos hoy? Una muy breve y simplificada revisión histórica…

Aunque los augurios de una supuesta singularidad parecen bastante nuevos en el horizonte de la AI, si vamos al fondo del asunto en realidad no lo son. Prácticamente desde el principio aparecen dos enfoques o teorías a saber: la denominada *AI fuerte*, que pretende que era posible crear máquinas con una inteligencia equivalente a la humana en todos sus aspectos y la *AI débil*, que -un poco más humilde- sólo pretende emular algunos aspectos de la inteligencia humana.

También podemos decir que existen dos grandes líneas: la simbólica y la no simbólica. La primera está inspirada en la lógica matemática, y dio lugar a los sistemas de reglas y representación de conocimiento. La segunda está inspirada en el funcionamiento del sistema nervioso y el aprendizaje en los organismos vivos. Los enfoques *biológicamente inspirados* están asociados principalmente con las ANN. En esta breve revisión histórica sólo nos referiremos a esta última rama de la AI que es la que ha logrado resultados y desempeños más impresionantes recientemente.

Los primeros modelos neuronales fueron propuestos por McCulloch y Pitts en 1943 a partir de lo que ellos llamaron *neurodos* y con los que pudieron recrear diferentes funciones lógicas a partir de su interconexión. Los neurodos funcionaban de forma muy similar a los modelos de neuronas artificiales actuales a partir de la suma de las entradas, ponderadas mediante pesos de conexión, que luego “pasaba” por una función de activación. Para ponerlo en contexto, recordemos que es recién en 1950 cuando Turing escribe su artículo donde presenta una extensa discusión acerca de la posibilidad de crear máquinas inteligentes y propone su famosa prueba. En 1957 Rosemblatt plantea su modelo de *perceptrón simple* y una regla de ajuste de los pesos de conexión para resolver problemas de clasificación sencillos (linealmente separables) en forma supervisada. Widrow y Hoff también plantean modelos similares con sus correspondientes reglas de aprendizaje de pesos (llamado ADALINE, del inglés ADAptative LINear Element). Al poco tiempo, un investigador del MIT llamado Marvin Minsky, desacredita estos trabajos mostrando que los perceptrones no eran capaces de resolver problemas triviales para la lógica digital como el del OR exclusivo (XOR). El trabajo de Minsky fue muy influyente y se lo ha identificado como una de las causas que frenaron el desarrollo de

las redes neuronales por varios años, por lo que se conoce a esta etapa como el *invierno de las ANN*. En los 80’s los trabajos de Rumelhart, Hinton y Williams proponen una forma de solucionar problemas de clasificación no linealmente separables como el del XOR a partir de la conexión de neuronas en capas (Perceptrón Multi-Capa, MLP) y de un método para encontrar los parámetros de la red llamado algoritmo de *retropropagación del error* (BP). En 1987 LeCun, en su tesis doctoral, retoma la idea de una variante de los MLP denominada *autocodificador* que utilizaba aprendizaje *semi-supervisado* para resolver problemas de codificación y compresión de imágenes. Durante los 90’s se multiplicaron las aplicaciones de los MLP a los más diversos campos y se vieron aparecer nuevas variantes que lidiaban mejor con imágenes (redes convolutivas o CNN) o señales temporales (memorias de corto y largo plazo o LSTM). También se demostraron las excelentes capacidades teóricas de aproximación de funciones de los MLP a la vez que manifestaron más claramente las limitaciones prácticas de los métodos de aprendizaje por los problemas de mínimos locales, sobre-ajuste o desvanecimiento del gradiente del error. Esto planteo un “techo” de desempeño que no podía ser superado, lo que generó una especie de *segundo invierno,* hasta que en 2006 Hinton aparece nuevamente proponiendo un “truco” que sería la base del *Aprendizaje Profundo* (DL). Profundizaremos, valga la redundancia, un poco sobre este tema hacia el final de esta charla. Este nuevo enfoque junto con el aumento de las cantidades de datos disponibles (Big Data) y las mejoras en el hardware de cómputo (GPUs y Clusters de PCs) produjeron un efecto potenciador en cascada que permitió sobrepasar los resultados del estado del arte en casi todos los campos de aplicación. El enfoque no tardó en aplicarse también a otras variantes de redes neuronales ya conocidas y combinado con otras técnicas clásicas como el aprendizaje por refuerzo permitió el diseño y prueba exitosa de sistemas de GAI, es decir que no utilizan información específica “a priori” acerca del dominio de aplicación. Ahí es donde estamos hoy.

Escenarios posibles para la mente de las maquinas

Vamos ahora a tratar de acercarnos a posibles respuestas a la pregunta original en los diferentes escenarios planteados, donde hemos tomado en particular a la consciencia como la característica de la mente humana de interés a emular u “obtener” mediante la AI.

Escenario 1: «aparición inesperada de la consciencia».

“¿Es posible construir una máquina dotada de inteligencia a nivel humano en el sentido analógico (mencionado en el apartado anterior), que resulte que tiene consciencia y subjetividad?”.

Hay varios partidarios de esta posición, de hecho, fue una de las primeras “hipótesis” de los partidarios de la AI fuerte. En este sentido han aparecido todo tipo de propuestas que podríamos resumir de la siguiente forma para cada caso[3]:

- **Fisicalismo:** el sustrato físico es suficiente, la consciencia es una propiedad física y todo ente material tiene algún grado de consciencia que varía con la complejidad del sistema (M. Kaku).
- **Biologicismo:** se requiere una base biológica y la consciencia surge a partir de algún tipo de interacción entre el cerebro y el cuerpo (S. Greenfield).
- **Funcionalismo:** lo importante es el procesamiento de la información y la memoria, si aumenta la complejidad algorítmica entonces en algún punto va a emerger la consciencia (Block, M. Minsky).
- **Comportacionismo**: si logra comportarse de forma indistinguible de una persona entonces es inteligente (Turing) y podemos suponer que tiene consciencia.

Desde las neurociencias este es un tópico de gran interés y aún abierto, donde se han propuesto diversas teorías para tratar de explicar los fenómenos que se observan en los humanos en distintas situaciones. Por ejemplo, durante la “suspensión” temporaria de la consciencia mientras dormimos o cuando sufrimos los efectos de una anestesia durante una cirugía. Sin embargo, desde las ciencias de la computación el interés principal que ha guiado a la comunidad ha sido más pragmático, tratando de lograr mejoras algorítmicas que impacten en métricas estándar de tareas conocidas pero difíciles.

Uno de los proyectos que podría ir en la dirección de lograr un mayor entendimiento de la estructura y funcionamiento del cerebro y a partir de allí orientar la búsqueda de propiedades emergentes es el denominado *Blue Brain* (2005). La idea consiste en reconstruir el cerebro pieza por pieza con el mayor detalle posible y construir un cerebro virtual en una supercomputadora. La potencia de cálculo necesaria es considerable: cada neurona simulada requiere el equivalente de una computadora portátil. Se han planteado diferentes etapas con metas claras en el futuro. El inconveniente con este enfoque es el *problema de la simulación* para proveer una consciencia y que está bien discutido en la comunidad (Tononi, Searle). Para comprender la limitación del mismo se suele usar una analogía: "Se puede simular una tormenta de forma casi perfecta en una computadora, pero de todos modos no hay forma de mojarse". Por lo tanto, aún en el caso de que fuera posible construir física o biológicamente una mente con consciencia, esta no podría aparecer en ninguna simulación computacional. Esto ocurre por ejemplo con las ANN, salvo

[3] Esto es sólo a modo de ejemplo de cada tipo y sin tener en cuenta todos los posibles matices.

que se construyan físicamente y respetando algunas propiedades y configuraciones particulares, como por ejemplo tener realimentación (Tononi).

Escenario 2: «creación de un zombi».

“¿Es posible construir una máquina indistinguible de un ser humano, pero sin consciencia, y por tanto sin perspectiva subjetiva?”.

Como ya se insinuó en el punto anterior, esto implicaría que se puede pasar la prueba de Turing sin necesidad de consciencia, realizando sólo una muy buena simulación o imitación. Una especie de zombis informáticos que actúan ejecutando reglas complejas pero sin ningún tipo de consciencia. Nuevamente hay muchos en la comunidad que comparten esta opinión (por ejemplo, Tononi, Searle).

Además, existe una comprobación práctica ya que varios algoritmos o programas han pasado la prueba de Turing con diferente grado de éxito, cada vez logrando mejores “imitaciones”, y en ningún caso sería razonable pensar que tuvieran consciencia. La última vez que ocurrió fue en 2014 cuando *Eugene Goostman*, un *chatter-bot* que interpretaba a un chico ucraniano de 13 años, logró engañar al 33% de los jueces del evento en un dialogo de 5 minutos, utilizando para ello varios “trucos” de los programadores. Si uno prolongaba un poco el interrogatorio resultaba más que evidente que se trataba de una imitación. Dadas las limitaciones de la prueba de Turing es que se han propuesto también diversas variantes y otras pruebas más difíciles. Por ejemplo, la del *Estudiante Universitario* que pide que el programa logre graduarse en la universidad para poder considerarlo como inteligente.

También hemos creado programas que resultan indistinguibles de los humanos (o incluso mejores) para resolver problemas o ejecutar tares complejas en algunos dominios. Aunque en general es claro que no tienen ningún tipo de consciencia se debe tener en cuenta que de todos modos pueden resultar dañinos o peligrosos para nosotros. Por ejemplo, tenemos hoy algoritmos que evalúan si somos aptos o no para tomar un crédito bancario y no es difícil imaginar las consecuencias de los errores de estos sistemas en la vida de una persona. También hay algoritmos que “diagnostican” mejor que un cardiólogo experto algunos tipos de patologías: ¿quién sería responsable en caso de una *mala praxis*? Sería aún más riesgoso si ponemos a un programa a controlar un sistema de misiles para la seguridad nacional de un país ya que no es necesario que *se dé cuenta* lo que está haciendo para que aplique ciertas reglas lógicas que resulten óptimas en algún sentido para cumplir su objetivo pero que puedan tener consecuencias nefastas. Aquí aparecen entonces varias cuestiones prácticas que es necesario considerar: ¿En qué condiciones pueden operar sin supervisión humana? ¿Cómo entreno los algoritmos sin sesgarlos con los ejemplos que le muestro?, entre muchas otras. Esto es lo que

ha motivado que comiencen a aparecer varios cursos de ética de la AI (por ejemplo, el de Joichi Ito del MIT Media Lab).

Escenario 3: «habilidades imposibles».

“¿Cabe la posibilidad de que el estar dotado o no de subjetividad implique necesariamente diferencias entre lo que hombres y máquinas pueden llegar a hacer?”.

Para tratar de dar una respuesta vamos a hipotetizar algunas habilidades que podríamos considerar relacionadas con una consciencia y vamos a ver si existe algún algoritmo o programa que muestre tener estas habilidades. La lista podría ser larga, aquí solo tomaremos algunos ejemplos: pintar, dibujar, criticar, hablar, tocar el piano, imaginar, escribir un libro, socializar, tener empatía, leer la mente.

Analicemos rápidamente cada una de ellas y algunos algoritmos disponibles:

3.1 Pintar/dibujar (cuadros): Gatys y colaboradores (2015) crearon un sistema basado en ANN que permite transformar una foto en un cuadro con un estilo artístico determinado. El estilo se puede aprender a partir de un conjunto de pinturas que se le muestran de un artista en particular. Por otra parte, Google ha desarrollado un nuevo estilo artístico denominado *Inceptionism* y que ya ha logrado vender muy bien varios cuadros en la conocida subasta de arte de San Francisco.

3.2 Criticar (obras de arte): Elgammal y Saleh (2015) desarrollaron un algoritmo que permite cuantificar la creatividad a partir de redes de arte. El mismo se entrenó con miles de pinturas de diferentes períodos de la historia y permite criticar y clasificar nuevas pinturas a partir de su comparación con pinturas similares.

3.3 Hablar/tocar el piano: Oord et al (2016) propusieron un modelo generativo para síntesis de voz y audio que modela la señal muestra a muestra partiendo de la representación temporal cruda utilizando de redes convolutivas. Los resultados de evaluación de la calidad del audio de la conversión de texto a voz (TTS) en pruebas subjetivas superan ampliamente a las del estado del arte. También se entrenó para generar música a partir de varias piezas

clásicas musicales de piano, lográndose nuevas interpretaciones con mucha naturalidad pero que conservan el estilo.

3.4 Imaginar: Zhang et. al. (2017) lograron desarrollar un método para la conversión de texto a imágenes con calidad foto-realista que utiliza redes antagónicas generativas apiladas. Este método logra generar imágenes de alta resolución a partir de la descripción textual de una escena y con muchos detalles, lo que podríamos considerar realmente como una capacidad de imaginación.

3.5 Escribir un libro: Max Deutsch (2016) entrenó una red tipo LSTM en los primeros cuatro libros de Harry Potter. Luego, le pidió que produjera un capítulo basado en lo que aprendió y los resultados son también muy convincentes. El algoritmo generó un capítulo completamente nuevo pero en sintonía con la historia y los personajes de la saga, sólo se realizó un ajuste de formato para ayudar a la legibilidad.

3.6 Socializar/tener empatía: Tay fue un robot de chat de AI lanzado por Microsoft a través de Twitter el 23 de marzo de 2016. Comenzó a publicar tweets ofensivos y racistas que obligaron a Microsoft a cerrar el servicio solo 16 horas después de su lanzamiento. Según Microsoft, esto fue causado por “trolls” que "atacaron" el servicio cuando el bot comenzó a generar respuestas basadas en sus interacciones con las personas en Twitter.

3.7 Leer la mente: Con Iván Gareis y otros colegas (2017) propusimos un método basado en autocodificadores ralos para estimar los promedios coherentes relacionados con la atención de eventos visuales. El método aprendió a reconocer patrones ruidosos en las señales cerebrales multicanal. De esta forma se pueden controlar dispositivos mediante la interpretación de las intenciones del usuario (interfaces cerebro-computadora o BCI).

Podría quedar como tarea la de buscar algoritmos para otras habilidades (tales como *creer*, *amar*, o *contar chistes*) pero creo que los ejemplos anteriores son suficientes para demostrar el tipo de capacidades que se han logrado y la dificultad para distinguirlas de las humanas en muchos casos. Claro está que para un ojo experto en cada dominio será más fácil realizar la distinción pero también es verdad que, de a poco, se van emulado cada vez más habilidades humanas y con mayor detalle. Por otra parte, nuestro análisis ha tomado cada habilidad por separado, como si fueran independientes, mientras que en nuestra consciencia todo esto se da de manera integrada, a partir de la interacción con otros seres humanos y el entorno, además de evolucionar constantemente teniendo en cuenta nuestra historia. Esto nos

da algunas pistas de donde podríamos buscar artefactos o *bugs* que nos permitan detectar las inteligencias sintéticas si la tecnología sigue desarrollándose[4].

Sobre la posibilidad *futura* de crear una inteligencia artificial con muchas de estas habilidades/características y que tenga consciencia las opiniones están divididas. Sin embargo, lo cierto hoy es que no se ha logrado demostrar que ninguna máquina sea consciente y de hecho estoy seguro que estos algoritmos no lo son. ¿Cómo puedo saberlo? Les hago una confesión: me siento como si fuera un mago profesional… ustedes me preguntan si la magia real existe entonces yo les muestro ejemplos de cosas espectaculares que "solo" podrían hacerse con magia real, pero la verdad es que yo tengo mis herramientas de mago y conozco los trucos. A continuación, se los voy a tratar de mostrar.

Mostrando algunos "trucos"

Para mostrar verdaderamente todos los trucos en detalle necesitaríamos un curso completo sobre AI, lo que queda fuera del alcance de esta charla. Es por eso que aquí solo vamos a describir algunas generalidades que permitan entender cuál ha sido la clave para mejorar el desempeño de las técnicas en algunos casos. Vamos a hablar primero del Aprendizaje Profundo y luego del impacto de este enfoque en el desarrollo de nuevas técnicas de GAI.

Aprendizaje Profundo

Ya en el segundo invierno de la AI se sabía que utilizar muchas capas de neuronas resultaba teóricamente más eficiente que pocas capas. Esto era también sugerido por la estructura de la corteza cerebral humana, pero no se tenían métodos prácticos de optimización que permitieran entrenar eficientemente estos modelos con muchas capas. Estas redes se componen de múltiples niveles de operaciones no lineales y la búsqueda en el espacio de los parámetros (pesos) es un problema de optimización muy difícil (mínimos locales, ruido, caos, inestabilidad, etc.).

El problema principal era que la estimación del gradiente del error en el algoritmo de retropropagación era peor a medida que se "alejaba" de la capa de salida. A este problema se lo denominó como *desvanecimiento del gradiente*. Simplificando bastante podríamos decir que lo que hizo Hinton (junto con algunos otros investigadores) fue demostrar que era posible descomponer el problema de entrenamiento multicapa supervisado en forma *voraz* en muchos sub-problemas de una sola capa pero no supervisados. Él comenzó trabajando con las

[4] Algo similar a lo que ocurría en la famosa película de ciencia ficción *Blade Runner* (1982).

denominadas *máquinas de Boltzman restringidas* (RBM), para luego armar capa por capa las *máquinas de creencias profundas* (DBM). Bengio hizo algo similar con los MLPs pero utilizando autocodificadores apilados y se encontraron relaciones y analogías importantes entre ambos enfoques. El resultado en general de este método no era óptimo, pero era mucho mejor que un modelo aleatorio, por lo que podía servir como un buen de punto de partida.

A este método se lo llamó *pre-entrenamiento* y se mostró experimentalmente que dejaba a la red relativamente cerca de un buen mínimo en la superficie de error aunque esta tuviera muchos mínimos locales. A partir de ese punto era posible realizar una etapa de *ajuste fino* de los pesos sobre la red completa mediante BP para acercarse más a ese buen mínimo. A partir de estas ideas se han propuesto nuevos algoritmos con mucho éxito relativo, derrotando en desempeño a las técnicas del estado del arte en muchas áreas, incluso en muchos casos con capacidades super-humanas como ya se ha mostrado.

Inteligencia artificial general

En la actualidad, los algoritmos de AI han sido diseñados, entrenados y optimizados por ingenieros humanos para lograr buenos resultados en una única tarea específica. En muchos casos superan a los humanos en habilidades, pero no pueden extender esas capacidades a nuevos dominios. Esto limita la reutilización, aumenta la cantidad de datos para entrenar y los deja sin generalidad ni desarrollo de “sentido común”. A este tipo de técnicas se las denomina como *AI Estrecha*. Como contrapartida, los propulsores de la GAI proponen que será capaz de superar estas limitaciones, aprender y proponer soluciones creativas para una amplia gama de tareas de múltiples dominios.

Lo interesante es que a partir del desarrollo de los métodos de aprendizaje profundo fue posible reutilizar viejos trucos potenciándolos para diseñar los primeros sistemas de este tipo con resultados prometedores en tareas muy difíciles. Este es el caso del aprendizaje por refuerzo, desarrollado hace varias décadas, y que se mezcló recientemente con DL. De esta forma AlphaGo (Google DeepMind) logró ganarle en 2016 un torneo a Lee Sedol quien había sido 18 veces campeón mundial de Go. Otro de los trucos utilizados fue que luego de aprender de diversas partidas entre humanos, se puso a jugar al programa contra copias de sí mismo, mejorando cada vez más su estrategia. AlphaGo ganó todos menos el cuarto juego y todos los juegos fueron ganados por resignación. La importancia de este hito es que el Go se considera un juego mucho más difícil que el Ajedrez porque es casi imposible planificar jugadas varios turnos hacia adelante por la explosión combinatoria. Es por ello que para ganar se requiere lograr cierta intuición o razonamiento aproximado como el que utilizan los jugadores humanos.

Conclusiones

En esta charla se han analizado diferentes escenarios en relación con la posibilidad de que a partir de los recientes desarrollos en la AI se pueda proveer de una mente a las máquinas. Los escenarios y el correspondiente análisis han surgido a partir de la interacción y el diálogo entre dos colegas de disciplinas diferentes: filosofía y ciencias de la computación. Se ha comenzado con una breve revisión de la historia de la AI, especialmente de los sistemas bioinspirados. Los recientes avances logrados en esta área se han debido a una serie de ingeniosos trucos que han permitido una mejora notable en los algoritmos y los métodos de optimización. A partir de estas técnicas y las grandes cantidades de datos disponibles es posible descubrir, modelar y analizar relaciones cada vez más complejas. Sin embargo, no asistimos todavía a un real cambio de fondo en la naturaleza ni en la concepción de los sistemas de AI. Es por ello que es posible afirmar que gran parte de las hipótesis acerca de que en el futuro cercano deberíamos esperar una singularidad en este sentido, tales como la aparición de una consciencia, no poseen todavía suficiente evidencia teórica o empírica para soportarlas. Creo que tal y como están planteadas se trata de *falsas profecías*, culmino con esta cita para seguir reflexionando:

> “Los ídolos de ellos son plata y oro, obra de manos de hombres. Tienen boca, mas no hablan; tienen ojos, mas no ven; orejas tienen, mas no oyen; tienen narices, mas no huelen; manos tienen, mas no palpan; tienen pies, mas no andan; no hablan con su garganta. Semejantes a ellos son los que los hacen, y cualquiera que confía en ellos.” (Salmos 115, 4-8)

Bibliografía

[1] W. S. McCulloch, W. Pitts, "A logical calculus of the ideas immanent in nervous activity", The bulletin of mathematical biophysics, Vol. 5, Issue 4, pp. 115-133 (1943).

[2] Turing, Alan, "Computing Machinery and Intelligence", Mind, Vol. LIX, Nº 236, pp. 433–460, (1950).

[3] Searle, John, “Minds, Brains and Programs”, Behavioral and Brain Sciences, Vol. 3, Nº 3, pp. 417–457 (1980).

[4] Christoph von der Malsburg “Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms” (1986).

[5] Lecun, Y., “Modeles connexionnistes de l'apprentissage (connectionist learning models)”. PhD thesis. Universite P. et M. Curie, Paris 6 (1987).

[6] Cybenko, G., "Approximations by superpositions of sigmoidal functions", Mathematics of Control, Signals, and Systems, 2(4), 303-314 (1989).

[7] Kurt Hornik, "Approximation Capabilities of Multilayer Feedforward Networks", Neural Networks, 4(2), 251–257 (1991).

[8] Mikel Olazaran, “A Sociological Study of the Official History of the Perceptrons Controversy”, Social Studies of Science, Vol. 26, No. 3, pp. 611-659 (1996).

[9] Minsky, Marvin, “The Emotion Machine: From Pain to Suffering”, Proc. of the ACM Conference on Creativity and Cognition, ACM Press (1999).

[10] Yoshua Bengio, “Learning Deep Architectures for AI”, Foundations and Trends in Machine Learning archive, Vol. 2 Issue 1, pp 1-127 (2009).

[11] Russell, S. and Norvig, P., “Artificial Intelligence: A modern approach”. Third Edition, Prentice Hall (2010).

[12] Edwin Chen, “Introduction to Restricted Boltzmann Machines” (2011).

[13] Marc' Aurelio Ranzato, “Neural Nets for Vision”, Tutorial on Deep Learning (2012).

[14] Interview with Eugene Goostman, the Fake Kid Who Passed the Turing Test, Time, June 9, (2014), http://time.com/2847900/eugene-goostman-turing-test/

[15] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, "Deep learning", Nature vol. 521, pp. 436–444 (2015).

[16] V. Mnih et al., “Human-level control through deep reinforcement learning”, Nature vol. 518, pp. 529–533 (2015).

[17] Leon A. Gatys, Alexander S. Ecker and Matthias Bethge, “A Neural Algorithm of Artistic Style” (2015).

[18] A. Elgammal and B. Saleh, “Quantifying Creativity in Art Networks”, 6th Int. Conf. on Computational Creativity (ICCC), USA (2015).

[19] Ian Goodfellow, Yoshua Bengio and Aaron Courville, “Deep Learning”, MIT Press, (2016) http://www.deeplearningbook.org

[20] Google’s ‘Inceptionism’ Art Sells Big at San Francisco Auction, Artnet News, (2016), https://news.artnet.com/market/google-inceptionism-art-sells-big-439352

[21] Aäron van den Oord et al, “Wavenet: a generative model for raw Audio”, Google DeepMind, London, UK (2016).

[22] Harry Potter: Written by Artificial Intelligence, Max Deutsch, (2016), https://medium.com/deep-writing/harry-potter-written-by-artificial-intelligence-8a9431803da6

[23] Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day, The Verge (2016), https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[24] D. Silver et al, “Mastering the game of Go without human knowledge”, Nature vol. 550, pp 354–359 (2017).

[25] Zhang et. al., “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”, ICCV (2017).

[26] I. Gareis, L. Vignolo, R. Spies and H. L. Rufiner, “Coherent Averaging Estimation Autoencoders applied to evoked potential processing”, Neurocomputing (2017).

[27] Tononi, Boly, Massimini and Koch, "Integrated information theory: from consciousness to its physical substrate". Nature Reviews Neuroscience. 17 (7): 450–461 (2017).

[28] The Ethics and Governance of Artificial Intelligence, Course MIT, Media Lab (2018), https://www.media.mit.edu/courses/the-ethics-and-governance-of-artificial-intelligence/

[29] Blue Brain Project, EPFL, Swiss (2018), https://bluebrain.epfl.ch/

[30] Is Consciousness Entirely Physical? INTERVIEW SERIES, (2018), https://www.closertotruth.com/series/consciousness-entirely-physical