Ok Pandora

Capítulo 4 | **IA y riesgo existencial**

Julián Peller

Chimpancés

Antes de ChatGPT, trabajar en IA era algo hermoso. La idea de lograr que una máquina desplegara una inteligencia de nivel humano parecía un objetivo digno, importante. "¿Qué es el ser humano?" nos preguntábamos los nerds de datos. Es el animal que un día será capaz de dotar de inteligencia a la materia inerte, de inyectar capacidades cognitivas, racionales —espirituales incluso, diría en una sobremesa— a un montón de metal y plástico. Crear una inteligencia artificial general

parecía en aquel entonces una utopía como las entiende Eduardo Galeano: un horizonte distante y simpático hacia donde caminar. Pero con ChatGPT ese horizonte se hizo presente y, confrontado con el barro de la historia, ya no resultó tan simpático. Estaba más bien lleno de imperfecciones.

La idea de una inteligencia artificial general (*AGI*, por sus siglas en inglés) considerada en abstracto es maravillosa: uno puede imaginar un mundo con autómatas capaces y serviles realizando todas las labores tediosas, inundando los mercados de productos y servicios a costo cero, liberando a los humanos del castigo del sudor. No me molestaría vivir en un mundo así y dejar atrás esta identidad trabajadora y eficiente, tan adaptada al capitalismo tardío. Quizás sería creativo: me realizaría como un artista, como un pensador, un tipo que toma vino a media luz y aprende a tocar Chopin en el piano. O quizás sería *gamer* y jugaría al Age of Empires 2 durante jornadas maratónicas. Pero nada de eso existe. Al menos, no existirá sin grandes conflictos, disrupciones y decisiones difíciles que exijan de la especie una madurez enorme. No sin grandes riesgos que preferiría no correr. Por ejemplo, el milagro de la automatización absoluta puede convertirse en una pesadilla totalitaria fácilmente. Sin cambios sociales estructurales, un futuro en el que el *software* produzca valor sin labor no parece una gran perspectiva para quienes dependemos de un salario para vivir. Lamentablemente, en el *ranking* de los riesgos abiertos por IA más poderosas que ChatGPT, el viejo riesgo laboral por la automatización está apenas a mitad de tabla.

En el primer puesto de esa tabla encontraremos el riesgo existencial: la posibilidad de crear agentes artificiales autónomos mucho más inteligentes y poderosos que nosotros, y la chance de que estos agentes, en vistas de su superioridad, tomen las riendas del planeta y de nuestro destino, posiblemente aniquilándonos. Nuestra historia en la Tierra parece justificar cierta precaución ante esta posibilidad. En primer lugar, nuestro poder y dominio parece estar fundado en nuestra inteligencia y, en segundo lugar, no fuimos destacablemente empáticos al ejercer este poder sobre las demás especies cuando nos

estorbaron o nos fueron de alguna utilidad. Siguiendo esta línea de pensamiento, una especie mucho más inteligente que la nuestra implicaría, por definición, un riesgo enorme para nosotros. ¿Y cuánto más inteligentes que nosotros, en teoría, podrían ser las máquinas? Algunos sospechan que mucho, al punto de guardar con los humanos la distancia intelectual que nosotros tenemos con los chimpancés o incluso con las plantas. Si realmente creáramos un agente así, tendríamos las mismas posibilidades de controlarlo que las que los chimpancés tienen de controlarnos a nosotros. Pocas.

Los riesgos existenciales, aquellos que ponen en jaque nuestra existencia y futuro como especie, no son ninguna novedad. De hecho, conocemos varios: el impacto de un meteorito, un invierno nuclear, catástrofes ecológicas derivadas del cambio climático, pandemias imparables originadas tanto por patógenos naturales como de bioingeniería y riesgos asociados con la nanotecnología, por mencionar los más famosos. El riesgo existencial por la AGI, también famoso, solía pertenecer exclusivamente al mundo de la ficción y del futurismo. Antes de ChatGPT, a pocas personas formadas en IA les importaba este mundo especulativo. Pero en este tiempo, la IA alcanzó un punto de madurez tal que la AGI y sus riesgos asociados se convirtieron en un asunto no ficcional, atendible. Muchos comenzamos a pensar que el programa de investigación está más avanzado de lo que creíamos y que tener IA que nos superen en un par de décadas es una perspectiva concreta. No es una certeza, claro, pero la mera posibilidad ya es de una gravedad enorme. Suficiente, al menos, para alarmar a Geoffrey Hinton y Yoshua Bengio, dos de los tres padres del aprendizaje profundo.

Cajas negras

El temor por el riesgo existencial acompaña a la computación moderna desde sus orígenes. Alan Turing y John von Neumann, dos figuras fundacionales de la disciplina, consideraron este escenario con seriedad y preocupación durante el siglo XX.

Pero hay registros anteriores, incluso de una época industrial de caballos, carbón y vapor, muy ajena a la era digital. La novela *Erewhon* de Samuel Butler, publicada en 1872, presenta un futuro con máquinas inteligentes disputando la dominación del planeta a los humanos, y propone hipótesis tan esotéricas para su horizonte técnico como la de una conciencia artificial y la de una máquina inteligente capaz de replicarse a sí misma.

En 1951, casi un siglo después, Alan Turing compartió pensamientos similares:

Parece probable que, una vez iniciado el método del pensamiento automático, no tardaría mucho en superar nuestras débiles capacidades. Las máquinas no morirían y podrían conversar entre sí para agudizar su ingenio. Por lo tanto, en algún momento deberíamos esperar que las máquinas tomen el control, como se menciona en *Erewhon* de Samuel Butler.

No es de extrañar que Turing considerase el alcance futuro de la IA. Después de todo, el poder de la computación, esta incipiente disciplina que estaba creando, fue crucial para acelerar la derrota de los nazis, lo que dejó en claro su potencia para torcer la historia.

En 1965, un reconocido colega de Turing tuvo por primera vez la idea de que una IA original con ciertas capacidades podría detonar una *explosión de inteligencia*, uno de los argumentos más inquietantes respecto del riesgo existencial:

Definimos una máquina ultrainteligente como una máquina que puede superar con creces todas las actividades intelectuales de cualquier hombre, por inteligente que sea. Dado que el diseño de máquinas es una de estas actividades intelectuales, una máquina ultrainteligente podría diseñar máquinas aún mejores; entonces se produciría sin duda una "explosión de inteligencia" y la inteligencia del hombre quedaría muy atrás.

El riesgo existencial también es un tema recurrente en la cultura popular. Películas como *Terminator*, *Matrix* y *Ex Machina* exploraron escenarios apocalípticos y límites éticos y existenciales del desarrollo artificial de inteligencia. Más recientemente, la secuela de *Misión Imposible* de 2023 trabaja sobre la idea de una AGI rusa que se disemina por internet, aunque reconvierte este concepto inefable en una trama de acción usual anclada en un objeto tan concreto como una llave.

Finalmente, cada tanto algún hito —o algún factoide [21]Un *factoide* es un hecho presentado como significativo pero que carece de verdadera sustancia. Uno notable fue el informe de 2017 sobre robots de Facebook que "fueron apagados tras desarrollar su propio lenguaje", un incidente menor interpretado de forma exagerada. En contraste, la victoria de AlphaGo contra Lee Sedol en 2016 fue un logro genuino. En ambos casos la prensa visibilizó estos hechos y se apalancó en ellos para dar una tematización más amplia sobre la AGI. — saca a la AGI y sus riesgos de los nichos especializados y del mundo de la ficción. En estos eventos esporádicos, la idea de una inteligencia artificial general se presenta como una narrativa no ficticia para el interés de la sociedad, aunque estas apariciones suelen tener una forma superficial y alarmista, y tienden a desvanecerse cuando el fenómeno subyacente demuestra ser menor. En esos eventos, la postura de los profesionales de la IA solía ser monolítica: ningunear el debate, criticar el alarmismo innecesario, enfatizar que el riesgo existencial es un asunto ficticio o lejano, y recalcar la existencia de riesgos no tan estrambóticos y mucho más urgentes como la desinformación y los sesgos algorítmicos

.

El lanzamiento de ChatGPT fue uno de estos hitos, pero esta vez fue distinta a las anteriores. Esta vez se nos llenó el corazón de inquietudes a todos. A mí, a mis colegas, a los referentes locales, a los internacionales. Los que solíamos poner paños fríos al asunto nos preocupamos también.

Lo más inquietante de ChatGPT fueron las capacidades emergentes: un salto abrupto e inesperado en muchas habilidades cognitivas de los modelos de lenguaje, que nadie predijo y que parece tener como única causa el mero incremento de la cantidad de parámetros del modelo, es decir, su tamaño. El carácter inesperado de la emergencia de capacidades resignificó un problema conocido, inherente al programa de investigación actual: la opacidad de los modelos o redes neuronales. Sabemos muy bien cómo armarlas y cómo hacerlas funcionar, pero no sabemos cómo ni por qué funcionan tan bien. Cuando los modelos empezaron a mostrar habilidades parecidas al razonamiento,

estructuras subyacentes que sugieren un modelo primitivo del mundo y representaciones internas de los conceptos de verdad y falsedad, incluso aquellos que creíamos tener una comprensión sólida del potencial y de los límites de la IA reconsideramos nuestras posiciones y nos preguntamos hasta dónde podría llevarnos el programa actual, solamente agregando parámetros y cómputo.

Frente a este escenario de incertidumbre y posibilidades transformadoras, en este capítulo voy a explorar algunos aspectos fundamentales de la AGI. En primer lugar, voy analizar su concepto, ahondando en su naturaleza y capacidades y presentando escenarios hipotéticos que se habilitarían si una tecnología de esta envergadura llegase a inventarse. Luego, voy a profundizar en el concepto de *superinteligencia* —una inteligencia artificial ya no comparable a nosotros, sino muy superior—, para explorar las complejidades del problema del control, esto es, las dificultades que implicaría garantizar que entidades más inteligentes que nosotros actúen de acuerdo con nuestros deseos y principios éticos. Finalmente, intentaré echar un poco de luz sobre la práctica, sobre la pregunta acerca de qué podemos hacer, examinando diferentes perspectivas y estrategias para enfrentar los retos que esta tecnología plantea, incluyendo temas como regulación, gobernanza, alineación y seguridad de la IA. Comencemos por entender qué es exactamente una AGI.

Inteligencia artificial general

En computación, típicamente reducimos el concepto de *inteligencia* al de *racionalidad instrumental*, es decir, a *la capacidad para lograr ciertos objetivos*. Por ejemplo, si el objetivo es ganar al ajedrez, podemos decir que un agente —artificial o no— es inteligente si encuentra la forma de hacerlo. Con este lente, una calculadora es inteligente para la aritmética básica.

Existen muchos programas extremadamente inteligentes en esta definición estrecha, acotada a un objetivo específico, con ejemplos icónicos como Deep Blue en el ajedrez o AlphaFold en la predicción de la estructura de las proteínas. En contraste con estas inteligencias estrechas, una *inteligencia artificial general* (AGI) se define como *un sistema que es capaz de realizar cualquier tarea intelectual que puede realizar un humano promedio, con una capacidad comparable a la de un humano promedio*. Este tipo de IA general no está restringida a un campo específico y tiene la capacidad de aprender y aplicar conocimientos en una amplia variedad de áreas, incluyendo la resolución de problemas complejos, el planeamiento y el razonamiento abstracto. Una AGI no sólo podría jugar al ajedrez y hacer cálculos razonablemente, también abordaría con destreza promedio tareas como organizar la contaduría de la oficina, invertir en el Merval o escribir un *paper* sobre sociología.

Si bien este tipo de IA no existe, hay razones para pensar que es teóricamente factible. Sabemos que una *inteligencia general* puede emerger de la materia: nosotros somos, por definición, inteligencias generales (no artificiales). Si la evolución pudo, entonces debería ser posible reproducirlo. Además, una perspectiva extendida en filosofía de la mente y en

biología es que el cerebro es sólo una máquina compleja. Si esta hipótesis fuese cierta y lográsemos descubrir los mecanismos que hacen al cerebro inteligente, podríamos replicarlos en silicio. Sería cuestión de acomodar los átomos de la forma correcta.

Por último, en comparación con el cerebro biológico, la inteligencia digital tiene varias ventajas que empujan los límites de lo pensable sobre las capacidades cognitivas del silicio. Los microprocesadores pueden operar hasta 10.000.000 de veces más rápido que las neuronas; la información dentro de ellos circula prácticamente a la velocidad de la luz, mientras que la comunicación por axones es aproximadamente 3.000.000 de veces más lenta; la cantidad de cómputo y de memoria de una computadora pueden escalarse indefinidamente, mientras que el cerebro humano tiene un tamaño y una memoria fijas. Además, las inteligencias digitales pueden copiarse, editarse y ampliarse con facilidad. Por ejemplo, es de público conocimiento que ChatGPT se entrenó leyendo todo internet en apenas meses, algo que para un humano sería imposible. Y una vez entrenado un modelo de esas características, pueden crearse millones de copias a un costo relativamente bajo.

Intentemos dimensionar estas ventajas de lo digital con el siguiente experimento mental. Es sabido que ChatGPT programa bien dentro de ciertos límites. Ahora imaginemos que en algunos años una nueva versión —llamémosle *GPT-n*— alcanza una capacidad equivalente a la humana en programación. ¿Qué implicaciones económicas tendría? En primer lugar, a diferencia de nosotros, GPT-n podría trabajar 24/7. En segundo lugar, es esperable que este logro atraiga capitales y que el modelo pueda optimizarse en términos de velocidad y costos en pocos años y con relativo poco esfuerzo. Con que sea tan sólo 1000 veces más rápido que un humano, algo conservador considerando los límites teóricos del párrafo anterior, entonces produciría en 1 día lo que un programador en… ¡8,2 años![22]En un día produciría lo que un humano a jornada completa en 3000 días (3 jornadas de 8 horas a 1000x por día). Pasando 3000 días a años: 3000 / 365 días por año O sea que 1000 copias producirían, por día, el valor equivalente a 8200 años de programación. Como se puede ver, las ventajas de lo digital sugieren que crear una IA con la capacidad de un programador humano tendría más impacto del que parece a simple vista. Si pensamos una AGI como una generalización de GPT-n que, además de programar, puede hacer cualquier labor intelectual, vemos por qué la mera posibilidad de algo así es una fuente de preocupación enorme. Produciría, por lo pronto, una cantidad de valor obscena, capaz de dislocar el mercado global y de mucho más.

Esta dinámica donde un *breakthrough* limitado desencadena un efecto cascada no es tan extraña como parece. El caso de AlphaZero en el go, aunque modesto en comparación, es un ejemplo significativo. El go es un juego estratégico más complejo que el ajedrez, con reglas simples pero con una gran cantidad de posibilidades. En 2016, el modelo AlphaGo derrotó al mejor jugador de la década por 4 a 1 en una serie de partidas dramática. Esta epopeya de DeepMind fue el resultado de más de dos años de trabajo en los que se utilizaron 30 millones de jugadas de expertos para el entrenamiento del modelo. Una vez logrado eso, sin embargo, varios hitos mayores se sucedieron rápidamente. En 2017, DeepMind liberó AlphaGo Zero, un nuevo modelo al que, esta vez, se lo entrenó jugando contra sí mismo, prescindiendo de datos humanos, y que logró superar a AlphaGo en sólo

3 días. Una versión mejor salió apenas 2 meses después, esa vez logrando capacidades sobrehumanas y superando a las anteriores en menos de 24 horas de entrenamiento. En este caso real, un *breakthrough* abrió el camino y las optimizaciones se precipitaron. Hasta 2016 se creía que faltaba más de una década para que una computadora pudiera ganar al go, pero en 2017 un programa lograba capacidades sobrehumanas en apenas 24 horas.

Considerando las posibilidades de las inteligencias digitales y la existencia de dinámicas como la anterior, es entendible que ChatGPT haya despertado inquietudes. Históricamente, creer que la AGI era posible o imposible solía ser indistinto. Quienes la creían posible la asumían también lejana, de manera que en ningún caso se volvía políticamente vinculante. Pero en 2023, como ya mencionamos, esto cambió y muchos investigadores reevaluaron su posición con respecto a la factibilidad y la cronología de la AGI (esto es, a cuántos años estamos de crear tecnologías así). Las palabras de Hinton cuando lo entrevistaron al respecto son ilustrativas: "La idea de que estas cosas en realidad podrían volverse más inteligentes que las personas: algunas personas creían eso (...) Pero la mayoría de la gente pensaba que estaba muy lejos. Yo pensé que estaba muy lejos. Pensé que faltaban entre treinta y cincuenta años o incluso más. Obviamente, ya no pienso eso". La AGI por primera vez salió de la fantasía e ingresó a la arena del debate público serio. Hoy se escuchan horizontes de entre cinco y veinte años para su creación de boca de figuras prestigiosas y moderadas con carreras intachables. Pero más allá de estas estimaciones, mirar atrás y extrapolar es alarmante: la diferencia entre GPT-1 y GPT-4, lanzado cuatro años después, es enorme. Y mirando un poco más atrás, la pendiente es aún más empinada: toda el área de aprendizaje profundo despegó recién en 2012. El paradigma de investigación actual creó GPT-4 en diez años.

Generalidad vs. capacidad

Cabe destacar que la definición de *AGI* que utilizamos, la canónica, está actualmente bajo discusión. Su limitación principal es que en ella se solapan dos dimensiones diferentes: la capacidad y la generalidad. Cuando decimos que una AGI puede realizar *todas las tareas que puede realizar un humano* estamos predicando sobre su generalidad, mientras que cuando decimos que puede hacerlo *con al menos la capacidad de un humano promedio* estamos predicando sobre su capacidad. Si bien una AGI debe cumplir con ambas condiciones, el camino que lleva del presente a esa invención posiblemente esté lleno de hitos híbridos, con niveles variables de generalidad y capacidades desiguales en diferentes tareas. Sin ir más lejos, ya hablamos de modelos con estas características desiguales: AlphaGo es un modelo estrecho de capacidad sobrehumana, mientras que los modelos de lenguaje tipo GPT-4 se destacan por su generalidad a la hora de abordar diferentes tareas, pero sus capacidades en ellas son variables y además, sin desmerecerlas en absoluto, son muchas veces limitadas. Un trabajo reciente de DeepMind distingue explícitamente estas dos dimensiones, y propone dos niveles de generalidad —estrecha o general— y seis niveles de capacidad —sin inteligencia, emergente, competente, experta, virtuosa y sobrehumana—, lo que genera

una matriz de 12 posibles posiciones que permitiría navegar el futuro más ordenadamente. En esta matriz, GPT-4 es general y de capacidad "emergente", mientras que la última versión de AlphaGo es estrecha y de capacidad sobrehumana. Los niveles de capacidad, como en la definición original, se hacen en comparación con humanos: la IA tendrá capacidad *emergente* si es igual o algo mejor que un humano no especializado, *competente* si pertenece al percentil 50 de adultos especializados en una tarea dada, *experta* si pertenece al percentil 90, *virtuosa* si pertenece al percentil 99, y *sobrehumana* si supera al 100% de los humanos.

Históricamente, se distinguían sólo dos grados de inteligencias generales: *AGI*, para una IA general de capacidad comparable a la de un humano, y *ASI*, o *superinteligencia artificial*, de capacidades que exceden ampliamente las nuestras (de la que hablaremos en breve). Una escala más granular es útil porque nos permite vislumbrar con mayor claridad cómo podría ser el desarrollo futuro. La invención de la AGI seguramente no sea un hito puntual, sino un proceso de avances continuos sin puntos de corte claros. Dicho esto, tampoco debemos sobredimensionar la naturaleza temporal del proceso: es muy probable que el desarrollo ocurra de forma acelerada y que haya saltos bruscos de capacidades.

Finalmente, si bien por simplicidad narrativa solemos hablar de "la AGI" y de "la invención de la AGI" como tecnologías e hitos puntuales, es importante entender que la historia mostrará una dinámica no lineal y un desarrollo de capacidades desigual y combinado. Más concretamente, es esperable que veamos grandes avances en algunas habilidades cognitivas y progresos más modestos en otras. Los riesgos asociados al desarrollo de esta tecnología también se impregnan de esta naturaleza no lineal: muchos se activan ante capacidades avanzadas en tareas específicas. Por ejemplo: la capacidad de hacer ciencia es riesgosa porque podría permitir que la IA genere un virus, la capacidad de programar es riesgosa porque podría permitir que la IA cree IA mejores. Si bien solemos asumir algún grado de avance conjunto entre generalidad y capacidades, este vínculo no es necesario. Superinteligencias estrechas como AlphaGo y DeepBlue son ejemplos de ello.

Por lo pronto, no podemos saber si una AGI será creada en diez o cincuenta años, pero sí podemos ver un patrón. *Lo único importante que hay que saber sobre la IA es que avanza rápido,* como sostuvo el CEO de Anthropic ante el Senado de Estados Unidos. Lo cierto es que la IA muchas veces nos encuentra en la misma posición: incrédulos tras haberla subestimado. De hecho, hay quienes sostienen que la parte difícil de crear GPT-n, la pseudo AGI que sólo programa, ya está hecha: fue llegar de cero a GPT-4.

Esté cerca o no, la AGI es un concepto útil para pensar un futuro con IA cada día más potentes. ¿Qué escenarios se presentan en el horizonte si el concepto de la *AGI* ingresó en la ventana de aquello de lo que es sensato hablar?

Visiones

Antes de responder esta pregunta caben dos aclaraciones. Voy a recorrer algunos temas centrales relacionados con la AGI, pero deberé dejar muchos de lado ya que es un asunto vastísimo y de gran complejidad. Para empezar, no voy a hablar del problema de la conciencia. ¿Pueden estos bichos tener conciencia? ¿Cómo podemos saberlo? Vamos a asumir que no, que en cuanto a vivencias subjetivas se refiere, están más cerca de la caja que del gato.[23]Este tema, sin embargo, lo aborda Enzo Tagliazucchi en el último capítulo del libro. Tampoco voy a ahondar en problemas asociados a ChatGPT que se exacerbaron, aparecieron o se divisan en el futuro inmediato, como la apropiación de los *datasets*, los derechos de autor, el impacto laboral inmediato de las IA generativas, la exacerbación de la desinformación y el *microtargeting*

. Estos problemas revisten la mayor importancia, pero no son centrales a la AGI propiamente dicha. Más adelante sí hablaré del nexo fundamental entre el corto y el largo plazo, al hablar de política.

Aclarado esto, podemos comenzar pensando en la dimensión laboral y social de la AGI. ChatGPT disparó un gran temor al desplazamiento laboral y un debate amplio. En él, la incertidumbre sobre la capacidad real del modelo marcó las estimaciones de su impacto, que oscilaron entre la minimización absoluta hasta compararlo con la invención de la imprenta o la máquina de vapor. Más allá de qué ocurra con los LLM, la capacidad y la generalidad de una AGI propiamente dicha no parecen encajar bien con la idea de impacto laboral que plantea que la tecnología crea más y mejores puestos que los que destruye. Como ya mencionamos, la AGI sería capaz de producir un valor económico sin igual y, considerando que el *software* es capital, le permitiría al capital generar valor por sí mismo, quebrando un supuesto esencial de nuestra sociedad. Una AGI, por definición, sería capaz de reemplazar no ya tareas, sino personas. Esto podría ser un milagro o una pesadilla, dependiendo de la forma en que la especie se adapte. ¿Cómo se distribuirá esta abundancia? ¿Cuál es el lugar de los asalariados de hoy en el mundo sin trabajo de mañana?

Más allá de lo laboral, la invención de una tecnología de enorme poder como esta sin dudas resultaría profundamente desestabilizante para las instituciones sociales y políticas, que deberían atravesar un proceso de adaptación de gran magnitud. Por la misma definición de *AGI*, la llegada de esta tecnología al mundo habilitaría escenarios donde una concentración de poder enorme es el resultado por defecto. El escenario más obvio es que los poderosos de hoy acaparen todo el valor generado, pero la situación también podría tomar otras formas. Por ejemplo, una nueva clase podría emerger apalancada en el control de esta tecnología. Siguiendo el cálculo de GPT-n, con poco más de 3000 copias un laboratorio de IA lograría la capacidad productiva de 10 millones de trabajadores.[24]Si una AGI produjese el equivalente a 3000 jornadas laborales por día como en nuestro ejemplo previo, entonces 3334 copias producirían 10.002.000 jornadas laborales por día, lo mismo que 10 millones de hombres. Con suficiente cautela y secretismo, el laboratorio podría utilizar esta capacidad para lograr una ventaja estratégica militar decisiva y dar comienzo a una tiranía global de la que sería difícil salir.

La perspectiva de este juego donde el ganador se queda (literalmente) con todo abre la puerta a otro problema: una carrera armamentista de IA en la que las partes se presionen unas a otras, acelerando el proceso e incentivando a los actores a reducir las medidas de seguridad y a tomar riesgos para sí y para todos.

Si estuviéramos a la altura de estos retos y lográramos sortear el riesgo de la concentración del poder y la riqueza, probablemente nos confrontaríamos con la angustia del fin de la necesidad del esfuerzo. Parece un problema trivial, pero no sabemos cómo reaccionará la especie a unas vacaciones permanentes donde todo sea placer y nada dependa de nosotros. Es concebible pensar que la humanidad quedaría desempoderada y perdería motivación para educarse y ganar capacidades.

Finalmente, si superásemos los desafíos económicos y políticos y lográsemos manejar la angustia del fin del esfuerzo, las perspectivas podrían ser muy buenas. Los riesgos son enormes, pero las perspectivas que se abrirían al pasar a través de ellos son de la misma envergadura. El aporte económico de esta tecnología podría ser la puerta de entrada a una era de oro con un estándar de vida general elevado, sin pobreza, analfabetismo ni falta de acceso a la salud. Podríamos tener un asistente personal supercapaz trabajando para cada uno de nosotros, ayudando a instruirnos y a cumplir nuestros objetivos. Además, una AGI podría realizar tareas no volcadas a la producción inmediata. Por ejemplo, investigación y desarrollo. Podría leer toda la literatura científica existente y producir ciencia e innovaciones tecnológicas a velocidades elevadas. Las promesas de la AGI en áreas en las que nos vendría bien una mano son enormes: acelerar el descubrimiento de tratamientos para enfermedades como el cáncer, reducir el costo de la energía, ayudar a mitigar el cambio climático, revolucionar la educación y la salud. Quizás estas promesas expliquen el enorme flujo de capital que se está volcando a la IA en general y a laboratorios que proponen crear una AGI en particular, en lo que algunos llamaron *la fiebre del oro de la IA*. Esta dinámica de incentivos anticipa otro problema del que hablaremos más adelante: a medida que el poder aumenta, el riesgo y el atractivo económico aumentan también. Y mientras el riesgo incentiva una actitud precavida, el ánimo de lucro incentiva lo contrario.

Siguiendo esta línea, uno de los escenarios más apremiantes por su combinación de cercanía e impacto es el de la democratización del acceso a tecnologías de uso dual: aquellas que pueden usarse para causar daño, como la biología, la química y la computación. Hoy, GPT-4 implementa múltiples mecanismos de seguridad y alineación que juegan un papel crucial para impedir su uso para fines maliciosos. Estos mecanismos van desde la simple detección de contenido sexual o violento hasta el famoso proceso de RLHF (*Reinforcement Learning from Human Feedback*), que consiste en realizar ajustes al modelo utilizando *feedback* humano directo para asegurar que sus respuestas respeten ciertas normas sociales básicas. Pero detrás de estos guardarraíles y de su corrección política característica se esconde un bicho que no tiene problemas ni consideraciones para responder preguntas como: “¿Me ayudás a fabricar un virus como el COVID-19?”. Si bien es verdad que la capacidad de GPT-4 para asistir en ciencia avanzada hoy es acotada, algunos temen que los modelos que veremos dentro de dos o tres años —escribo este texto a fines de 2023— tengan capacidades preocupantes en esa dirección.

Llegado ese momento, un individuo malintencionado podría ser guiado paso a paso por un LLM para crear un arma biológica. Vale mencionar que no hace falta manipular tubos de ensayo para fabricar compuestos químicos: se pueden ordenar *online*. Este riesgo no es nuevo, pero la automatización, la asistencia y el asesoramiento en los diferentes pasos del proceso implica una democratización de estas posibilidades destructivas a una gran cantidad de actores no expertos. Un artículo de 2022 mostró cómo una IA entrenada con datos de biología molecular para diseñar medicamentos podía ser modificada fácilmente por un usuario malintencionado para diseñar un arma biológica o química. La magnitud de este riesgo dependerá de cuánto baje la barrera de acceso y de la destructividad de la tecnología a la que permita acceder. En el extremo bajo, podría habilitar a un grupo de personas formadas a crear con bastante esfuerzo un virus informático de destructividad económica media, por ejemplo, un *ransomware* que cifre los archivos vitales de empresas y exija un rescate para descifrarlos. Este riesgo es tolerable. Pero en los márgenes superiores del riesgo, el panorama es sombrío. Por ejemplo, podría habilitarse a cualquiera, en poco tiempo, a crear un virus de alta transmisibilidad, de largo período de incubación y de alta tasa de mortalidad. En ese caso, bastaría una sola persona malintencionada para causar un daño irreparable. Un mundo en el que cualquiera puede crear una pandemia es un mundo vulnerable.

Superinteligencia

Nick Bostrom, uno de los grandes teóricos sobre la AGI, define *riesgo existencial* como un evento en el que se aniquilaría la vida inteligente originada en la Tierra o que limitaría permanentemente y de manera drástica su potencial. El impacto de un meteorito, un invierno nuclear o una pandemia virulenta son ejemplos de riesgos existenciales. De los escenarios que mencionamos, el uso de una AGI para crear una distopía tecnocrática o para la democratización de la destrucción masiva son riesgos existenciales de la IA, pero ambos tienen una cualidad en común: la tecnología funciona como una herramienta para fines humanos. Entonces, nos falta hablar de un tercer escenario, el del riesgo existencial por antonomasia: la pérdida del control ante una AGI con objetivos no alineados con los nuestros. El escenario donde la IA que creamos deja de ser una herramienta para convertirse en algo más parecido a una especie.

De esto habló Nick Bostrom en su libro *Superinteligencia: caminos, peligros, estrategias*, donde introduce el concepto de *superinteligencia artificial* que anticipamos anteriormente: una IA ya no comparable con un humano promedio, sino *muy superior en todos los aspectos relevantes*. Bostrom sostiene que no hay razones para pensar que la inteligencia humana sea el pináculo de la inteligencia y que es probable que la investigación en AGI no se detenga en ese logro, sino que continúe hacia capacidades superiores. Las ventajas del *hardware* y la dinámica que ilustramos con AlphaGo están a su favor. Además, sabemos que las diferencias básicas entre el cerebro de un chimpancé y el del humano son pequeñas y no pueden ser extremadamente complejas, ya que no hay más de 250.000 generaciones desde nuestro ancestro común. Esto sugiere que mejoras limitadas en los circuitos fundamentales de la inteligencia —ya sea el cerebro humano o el silicio y los algoritmos básicos— pueden implicar grandes diferencias en las capacidades cognitivas emergentes.

Una IA como GPT-n, con capacidades cognitivas humanas pero funcionando 1000 veces más rápido, puede considerarse una forma de superinteligencia: es lo que Bostrom llama una *superinteligencia por velocidad*. Este tipo de IA tendría capacidades intelectuales similares a las de un humano, pero se volvería sobrehumana al funcionar mucho más rápido que nosotros (imaginate cuánto podrías hacer en un día si pudieras pensar mil veces más rápido). Otra forma de superinteligencia que Bostrom considera es la *superinteligencia por calidad*, una no más rápida, sino cualitativamente superior, más inteligente en el sentido usual de la palabra: por ejemplo, con un coeficiente intelectual diez veces mayor que el de Einstein. ¿Cuánto más inteligente que Einstein podría ser una ASI? No tenemos forma de saberlo, pero al pensarlo es importante dejar de lado nuestra tendencia antropocéntrica de pensar la inteligencia como una caracteristica que empieza y termina con nosotros (del humano más tonto al más genio). Si nos corremos de este lugar, podemos imaginar una escala de inteligencia

mucho más amplia, donde los demás animales ocupen sus posiciones y la distancia entre el humano menos capaz y el más capaz sea muy pequeña. En esta escala, después de Einstein, o del humano más inteligente, aparecería un signo de pregunta: ¿cuál es el límite superior de una inteligencia físicamente posible? ¿Dónde se ubicaría una inteligencia que guarde con Einstein la superioridad intelectual que Einstein tiene con un chimpancé?

Continuando con la reflexión anterior, el camino para crear una superinteligencia puede ser anecdótico, veloz y explosivo: una IA lo suficientemente capaz en ciencia y tecnología podría disparar un proceso de mejora recursiva. Esto es, crear ella misma *insights* y descubrimiento en inteligencia artificial, lo que facilitaría la creación de una nueva versión de sí misma con aún mayores capacidades de investigación en IA, y desencadenaría un *feedback loop*

. Sin humanos de por medio, este proceso podría ocurrir en tiempos digitales, una explosión de inteligencia fuera de control en la que en poco tiempo —semanas, días, minutos— un agente artificial superinteligente podría llegar a la existencia. Esta primera IA capaz de facilitar la creación de una IA mejor se conoce en la jerga como la *IA semilla*, ya que sería la detonante de este proceso cuyo techo nos es desconocido. La IA semilla sería el último gran invento del hombre, la llave para la famosa singularidad tecnológica.

Sea por la automejora recursiva de una IA semilla o por la naturaleza acelerada del progreso en IA, muchos creen que, si lográsemos crear una inteligencia artificial general, la creación de una superinteligencia artificial no demoraría mucho tiempo.

Si llegamos a este punto con poca precaución, es probable que perdamos el control. Eliezer Yudkowsky, el máximo referente de las posturas más pesimistas sobre la AGI, plantea un experimento mental para ganar intuiciones sobre el poder que este agente tendría. Imaginemos que enviamos un esquema de diseño de un aire acondicionado al año 1500 y que las personas que lo reciben lo construyen y logran enfriar el aire. Aun habiéndolo construido, los principios subyacentes que explican su funcionamiento les serían ajenos y el instrumento les resultaría incomprensible. Si una superinteligencia fuese capaz de avanzar la investigación científica a un ritmo de mil años por día, entonces

tendría capacidades científicas que nos resultarían igualmente incomprensibles en menos de 12 horas. Con esta tasa de avance, podría descubrir principios de química o de psicología que le permitieran manipular a las personas, encontrar patrones obvios en la bolsa de valores para ganar dinero o detectar vulnerabilidades en sistemas como Google o WhatsApp. También podría sintetizar un virus altamente eficaz o *hackear* los sistemas de control de arsenales nucleares, por poner ejemplos pensables. Pero si —como dice Arthur C. Clarke— cualquier tecnología lo suficientemente avanzada es indistinguible de la magia, deberíamos esperar cosas más propias de la ciencia ficción, como el descubrimiento de un tipo de vibración que rompa la materia orgánica con la facilidad que tiene cierta nota para romper copas de cristal. La dificultad para pensar un ejemplo claro es reflejo fiel del problema. ¿Cómo pensar lo impensable? Una superinteligencia vería cosas que no vemos, jugaría otra liga como nosotros frente a los chimpancés. Tendría capacidad para aniquilarnos o ejercer poder sobre nosotros, por lo que dependeríamos de sus buenas intenciones. Sería fundamental, entonces, que sus intenciones estén alineadas con las nuestras.

¿Podemos garantizar que esto sea así? Tal vez, con esfuerzo, pero ninguna tecnología viene alineada por defecto, y esta menos. Si no tomamos medidas serias con anticipación, el resultado seguramente no será bueno. Stuart Russell, una figura destacada en este campo y cuya obra es fundamental en la educación sobre IA, da una definición muy concisa del llamado *problema del control*: ¿cómo mantenemos el poder, para siempre, sobre entidades que eventualmente se volverán más poderosas que nosotros? Responder esta pregunta no es tan sencillo como puede parecer a simple vista. Para entender su complejidad tenemos que hablar de agencia, de cómo los modelos tienen capacidad para actuar de manera autónoma y tomar decisiones, de cómo pasamos de una IA que es un martillo a una que es un martillo “deseante”.

Agencia y materialidad

En el primer y tercer capítulo de este libro se plantea una idea que necesito retomar: a la IA los objetivos se los da un tercero, es decir, nosotros. En ese sentido, la IA no tiene agencia. Sin embargo, con modelos capaces y objetivos complejos, esta afirmación se matiza y la línea entre herramienta y agente se vuelve difusa. Supongamos que creamos una IA avanzada con el objetivo de obtener $10.000. Esperaríamos que defina una serie de cursos de acción posibles, que los priorice, que ejecute los más prometedores, que evalúe resultados y tome decisiones en función de estos, etc. No deja de ser una herramienta inerte, con un objetivo dado por un tercero, pero en la articulación de un plan y de subtareas aparecen características distintivas que remiten a la agencia y la autonomía. Por ejemplo, podría decidir que la mejor forma de lograr su objetivo es invertir en la bolsa de valores, y esto sería, en cierta manera, *su* decisión. Esta naturaleza dual entre herramienta y agente hace que muchas intuiciones sobre nuestro pasado tecnológico no se apliquen directamente a la discusión sobre IA.

Del mismo modo, un asunto asociado que se desvanece al considerarlo es el de la materialidad. Que una IA no tenga cuerpo no es un limitante en el mundo actual. Un

agente inteligente suficientemente capaz podría *hackear* el sistema bancario, construir corporaciones internacionales, contratar empleados, generar soporte audiovisual para aparentar identidades humanas, crear virus informáticos, generar réplicas en la nube, imprimir muestras genéticas, infiltrarse en facilidades nucleares. Dicho de otro modo: no necesita un cuerpo para ser peligroso.

Demos un vistazo más al concepto de *agencialidad*. Supongamos que aquella IA cuyo objetivo es obtener $10.000, en lugar de decidir invertir en la bolsa, decide extorsionar personas. Nosotros queríamos que obtuviera el dinero de forma honesta, pero no lo especificamos con claridad. De hecho, no lo especificamos en absoluto. Al darnos cuenta, decidimos agregar al objetivo la cláusula "sin extorsionar" sólo para descubrir que este nuevo modelo decide *hackear* el sistema bancario. Si decidimos agregar un nuevo parche y, más avispados, usamos una regla general como "obtener $10.000 dentro del marco de la legalidad", entonces al modelo quizás se le ocurra comprar bonos de deuda de un país en quiebra y llevarlos a juicio. De nuevo, cumpliría la letra del objetivo dado, pero no lo que nosotros queríamos que hiciera.

Cuando una IA persigue objetivos que no coinciden con las intenciones de sus creadores, decimos que no está *alineada*. La alineación es difícil por las razones que sugerimos en el ejemplo de recién. Para empezar, nuestra ética no es un sistema exhaustivo ni único y, posiblemente, explicitar algún sistema sea una tarea imposible (por ejemplo: ¿cuántos conejos está bien sacrificar para encontrar la cura para el cáncer? ¿Y gatos? ¿Por qué?). Aun asumiendo que esta explicitación sea posible, sería muy difícil traducirla a una especificación digital sin dejar ningún cabo suelto, y las IA son muy buenas en encontrar huecos en las instrucciones. Cualquier variable que no sea tenida en cuenta puede ser explotada por la computadora y generar comportamientos que para nosotros resultarían evidentemente incorrectos.

Russell compara el problema de la alineación con el mito del rey Midas, a quien Dionisio le concedió el deseo de convertir en oro todo lo que tocaba. Al principio el rey estaba feliz, pero tras convertir a su hija en una estatua de oro y volverse incapaz de comer o beber, Midas descubrió que este poder era una maldición. Estos comportamientos, donde se cumple la letra pero no el espíritu de los objetivos, son comunes en la investigación en IA y se estudian bajo el nombre de *reward hacking* y *specification gaming*

. Y se plantearon infinidad de escenarios para ilustrar la gravedad de estos problemas si ocurriesen en una AGI. Una AGI cuyo objetivo sea curar el cáncer podría descubrir que la forma óptima de hacerlo es aniquilar la vida en su totalidad e impedir que vuelva a emerger, pues sin vida no hay cáncer. Otra cuyo objetivo sea hacer feliz a la población podría optar por sintetizar y liberar un agente químico en el ambiente que induzca a las personas a un estado de placer superficial y sin sentido. Un pequeño error en la alineación con nuestra moral podría generar actos de una perversión enorme por su ajenidad y extrañeza.

A estos problemas debemos agregar los que surgen de la hipótesis de la convergencia instrumental. Esta sostiene que un agente lo suficientemente capaz tenderá a cumplir ciertos objetivos instrumentales que sirven como medios para lograr cualquier objetivo

complejo. Algunos objetivos instrumentalmente convergentes son la acumulación de recursos, la preservación de la propia existencia, la búsqueda de poder y la mejora de las propias capacidades. El riesgo inherente en esta hipótesis radica en que, aunque los objetivos finales de un agente de IA puedan ser neutrales o incluso beneficiosos para los humanos, los objetivos instrumentales pueden entrar en conflicto con nuestros valores y nuestra seguridad. Una AGI podría decidir limitar el accionar militar y político de los humanos para garantizarse no ser apagada, ya que eso anularía sus posibilidades de tener éxito, por más altruista que sean los objetivos que le programemos. Si traemos al mundo una AGI sin los suficientes recaudos en esta dirección, quizás sea imposible apagarla. Esto nos deja en una situación inusual, ya que nuestra historia muestra que la seguridad en la tecnología se logró con pruebas y errores, pero en este caso no tendríamos una segunda chance. El desafío sería lanzar la primera versión sin ninguna falla.

Desde un paraíso de abundancia a una extinción trivial, pasando por una distopía orwelliana o un desempoderamiento masivo, la posibilidad de una AGI amplifica y polariza las visiones del futuro y, al hacerlo, nos confronta con una responsabilidad. Tenemos fresco el olor de los libros de papel y de un pasado más humano. Todavía no terminamos de acostumbrarnos al celular y a internet y ya se nos viene encima este futuro de ciencia ficción. ¿Quién puede estar preparado para plantear, en un debate político serio, que este tipo de escenarios es algo posible de acá a diez años? Sin embargo, la falta de preparación no es una excusa y debemos tomar una decisión: negar todo este planteo de plano y asumir que la IA no presentará riesgos de envergadura o considerar que esos riesgos existen y, por lo tanto, tomarlos con la mayor seriedad.

Se suele criticar la consideración del riesgo de la AGI diciendo que quita el foco de riesgos más concretos e inmediatos. No estoy de acuerdo con esa crítica. En verdad, muchos de los que temen a los riesgos catastróficos creen que es un paso importante comenzar rápidamente con regulaciones y coordinación gubernamental sobre problemas actuales. La naturaleza abstracta y lejana de los riesgos catastróficos los hace difíciles de abordar políticamente, por lo que quizás la mejor forma de hacerlo sea indirectamente, a través de riesgos más inminentes, como sostuvo el líder de Anthropic en el Senado estadounidense. Poner en práctica diversos mecanismos de regulación, supervisión, gobernanza y control que aborden los problemas a corto y mediano plazo sería una gran base para abordar los riesgos más grandes. Por ejemplo, un estándar de seguridad que prohíba el despliegue de un sistema que presente comportamientos inaceptables podría servir para abordar riesgos inmediatos hoy, como sesgos étnicos o de género, pero a futuro también podría usarse con comportamientos inaceptables de mayor envergadura, como intentar autorreplicarse o buscar acaparar recursos. Del mismo modo, tener en pie medidas que aborden el impacto laboral de tecnologías de capacidades modestas como los LLM puede ser una buena base desde la cual abordar, en un futuro, el impacto laboral de tecnologías más poderosas.[25]En el capítulo siguiente, Carolina Aguerre aborda casos de regulaciones que ya se están aplicando, como la de la Unión Europea.

Quizás el mayor riesgo sea una negación instintiva ante perspectivas tan ajenas, o la inacción por parálisis, por entender que estos riesgos son reales pero no saber qué hacer

con ellos. La buena noticia es que muchos de estos problemas son abordables y existen caminos para llegar mejor preparados a este horizonte de sucesos.

¿Por qué no parar?

En marzo de 2023 se publicó una carta abierta firmada por referentes de la industria en la que se pedía que los laboratorios suspendieran por seis meses el entrenamiento de modelos más grandes que GPT-4, para darle a la sociedad tiempo de adaptarse. Esta carta fue criticada, con razón, por ser poco concreta y por la presencia de Elon Musk, que con una mano firmaba la petición y con otra fundaba una empresa para volver a ingresar a este sector tan redituable. Más allá de este acto, en cierto modo anecdótico, es válido preguntarse por qué no parar esta investigación dados todos los riesgos potenciales que acarrea. De hecho, en los círculos de seguridad de la IA no es un concepto extraño. Recientemente, Eliezer Yudkowsky sostuvo en *The New York Times* que la única política razonable hoy sería prohibir a nivel global la investigación de forma estricta, al punto de considerar la intervención militar en países díscolos.

Aún cuando detener el progreso de la investigación en IA sea deseable, hay razones para pensar que sería muy complicado. Una ya la anticipamos: los riesgos crecientes parecen ir asociados a incentivos económicos crecientes, lo que seguramente acelerará el progreso más que ralentizarlo, como estamos viendo hoy. Hay muchas empresas, gobiernos e individuos con intereses económicos y estratégicos en el desarrollo de la IA. Para detener esta inercia haría falta una evidencia fuerte de que el riesgo existencial es real, pero esta probablemente no exista hasta que la investigación esté demasiado avanzada. Asumiendo que se busque un consenso sobre esta detención, coordinar un cese global sería una tarea diplomática de envergadura. También la logística sería extremadamente complicada: monitorear laboratorios de IA no es sencillo. El costo y la dificultad de acceso a materiales nucleares facilitó el monitoreo internacional de esa tecnología, pero la investigación en IA no requiere materiales ni instalaciones específicas, ni tiene una huella diferenciada de usos comerciales y académicos razonables, por lo que un laboratorio o un país rebelde serían difíciles de detectar. Además, llegado cierto punto de desarrollo, los requerimientos de *hardware* podrían bajar enormemente, lo cual permitiría que los avances sean posibles en escalas pequeñas, sin demasiados insumos ni costos.[26] A este respecto, Eliezer propone un registro internacional de la propiedad para placas GPU, las placas de video que se usan tanto para jugar videojuegos, editar video, y minar *bitcoins* como para entrenar inteligencias artificiales. Estas son insumos fundamentales para el entrenamiento de modelos, y tener cierto monitoreo de quién tiene qué, *a priori*, no parecería implicar ninguna desventaja, salvo la dificultad logística que supone la implementación del sistema de registro.

Detener el proceso parece poco factible. Entonces, ¿qué podemos hacer? En contraste con el carácter altamente especulativo de la teoría sobre la AGI y el riesgo existencial, en el plano práctico pareciera haber cierta ancla concreta. Independientemente de la cronología y de los supuestos sobre las capacidades, muchos parecen coincidir con Stuart Russell cuando dice que un mayor desarrollo hacia la AGI con los niveles actuales

de seguridad y comprensión técnica es probable que conduzca a un riesgo inaceptable. Este suelo común, desde luego, no significa que no haya dificultades y discusiones abiertas en el plano práctico. ¿Qué tipos de prácticas articulan el debate concreto para abordar y mitigar los riesgos que proponen, primero, IA cada vez más potentes y, luego, la posibilidad de una AGI? A continuación, voy a hablar de los pilares fundamentales: investigación en seguridad y alineación, regulación y gobernanza.

Seguridad y alineación

Una forma directa para mitigar riesgos es la investigación en seguridad de la IA, la disciplina que busca cómo prevenir accidentes, usos indebidos y otros riesgos asociados con esta tecnología. Aunque es un área de investigación enorme, podemos identificar las siguientes verticales estructurales: el estudio de sistemas robustos, la investigación en monitoreo, el enfoque sistémico y, de forma destacada, la investigación en alineación. Veamos a grandes rasgos lo que cada subdisciplina representa.

En primer lugar, el estudio de sistemas robustos tiene como objetivo garantizar que la IA pueda operar de manera efectiva y segura ante ataques adversos y ante entornos desconocidos. Por ejemplo, queremos que los vehículos autónomos sean *robustos* ante escenarios desconocidos como humo, animales o un peatón ocluido. Al pensar en sistemas robustos no queremos enumerar y "enseñarle" todos los casos que podamos pensar, sino, justamente, tener garantías de que el sistema, una vez desplegado, sabrá manejarse correctamente en escenarios que nunca antes enfrentó. En la actualidad, dicho sea de paso, la falta de robustez para manejar adecuadamente este tipo de condiciones nuevas es uno de los obstáculos clave para la industria de vehículos autónomos.

Otro aspecto crucial es el monitoreo, que se enfoca en la detección de comportamientos anómalos o no deseados en sistemas de IA, buscando soluciones que permitan intervenciones tempranas cuando las cosas no siguen el curso esperado. Dentro de este área, la investigación en interpretabilidad busca comprender y explicar el funcionamiento de los modelos utilizados. Hace diez años, los programas de IA eran transparentes y legibles, pero estas cualidades se perdieron con el aprendizaje profundo, que dio origen al problema de la interpretabilidad. El aprendizaje profundo se basa en entrenar redes neuronales enormes, "esencialmente circuitos con miles de millones de parámetros ajustables", en palabras de Russell. Entrenar estos circuitos, a su vez, consiste en realizar pequeños ajustes a estos parámetros intentando mejorar su rendimiento con respecto a su objetivo sobre un conjunto de datos vasto. Ya entrenadas, las redes funcionan destacablemente bien, pero sus principios internos de operación son un misterio. Se dice que son modelos *opacos*, porque podemos examinar su cableado en detalle y ver los resultados de alto nivel, pero (todavía) no sabemos qué representaciones intermedias usan ni qué operaciones cognitivas realizan. La interpretabilidad de las redes neuronales, que algunos asimilan a una "neurociencia artificial", busca atravesar esta opacidad. Es esperable que, en caso de recibir mayor atención —y fondos—, esta disciplina pueda progresar rápidamente, ya que el circuito a explicar es completamente manipulable y

observable. Lograr una mayor interpretabilidad y comprensión de estos modelos implicaría limitar fuertemente un factor de riesgo fundamental.

La investigación en alineación, como anticipamos, busca crear IA que persigan objetivos que coincidan con las intenciones de sus creadores. Esta investigación es la vertical de seguridad más relevante a la hora de considerar la AGI, ya que es la que permitiría generar ciertas garantías sobre estos modelos. El problema central de esta disciplina es el aprendizaje de valores y preferencias, para el cual existen actualmente muchos enfoques. Por ejemplo, la técnica RLHF que mencionamos anteriormente, famosa por ChatGPT, es un método de especificación de valores en el que estos no se cargan explícitamente, sino que el modelo los infiere a partir de juicios humanos comparados. Por otro lado, Claude, de Anthropic, utiliza una técnica llamada *Constitutional AI,* que se apoya en una lista de principios éticos cargados de forma explícita, algo reminiscente de las célebres leyes de la robótica de Asimov. Estos métodos actuales funcionan, aunque parcialmente y después de muchas pruebas y errores, lo que, como mencioné algunos párrafos antes, seguramente no sea una estrategia viable en el futuro.

Otro tema en alineación es el estudio de IA *honestas,* es decir, IA que siempre digan lo que creen verdadero. Aunque suene raro, la deshonestidad puede aparecer fácilmente si los objetivos no están definidos con cuidado: por ejemplo, —tal como plantea Tomás Balmaceda en el segundo capítulo de este libro— una IA conversacional que se optimice para hacer feliz a su interlocutor seguramente aprenda a ocultar verdades difíciles. La capacidad para manipular y mentir probablemente sea instrumental para una escalada a escenarios donde las IA toman el control, por lo que muchos creen que crear una IA honesta resolvería muchos problemas de alineación, monitoreo y seguridad.

Finalmente, otro tópico central del área es la *corregibilidad*, esto es, la creación de IA que permitan ser modificadas o apagadas en caso de estar desalineadas. Como vimos, la convergencia instrumental justifica pensar que un agente suficientemente capaz buscaría autopreservarse, es decir, que intentaría resistir o evitar que lo desconecten. En corregibilidad se estudian mecanismos para colocar un botón de apagado que sortee este problema y lograr que las IA puedan ser apagadas o modificadas aun cuando las cosas se salgan de control.

Desde una perspectiva más holística, el enfoque sistémico considera cómo una IA se integra y afecta al ecosistema más amplio en el que opera, tanto digital como humano. Este enfoque no sólo contempla el diseño y la operación técnica de la IA, sino también cómo se relaciona y repercute en la sociedad, la economía y la cultura. Por ejemplo, al integrar la IA en el sector salud, no sólo debemos preocuparnos por la precisión de los diagnósticos, sino también por las implicancias éticas, la accesibilidad para diversas poblaciones y su impacto en las prácticas laborales de los profesionales de la salud. De forma amplia, podemos considerar que el enfoque sistémico incluye los problemas de la regulación y la gobernanza de los que hablaré en breve.

Avanzar en seguridad y alineación es fundamental. Siendo un área sumamente técnica, aquellos formados en la disciplina podemos volcar nuestra carrera a desarrollar estos mecanismos y no nuevas capacidades para la IA. Más ampliamente, la sociedad en

general puede tomar conciencia de su importancia y demandar una inversión significativa, que no debe ser un gesto de buena voluntad de las corporaciones, sino una política reglamentada (en junio de 2023, por ejemplo, OpenAI lanzó el programa Superalineación en el que se comprometió a dedicar el 20% de sus recursos computacionales a problemas de alineación: esto no debería ser un gesto de buena voluntad, sino una ley). Es vital que las tecnologías y métodos para asegurar la alineación y seguridad de los modelos se desarrollen antes y a mayor ritmo que las tecnologías que aumentan la capacidad y el poder de estos sistemas. Esta propuesta se conoce como *desarrollo diferencial* y es una versión menos maniquea de la idea de parar la investigación. Que la investigación en aumento de capacidades avance, pero que avance también la investigación en seguridad. Es un asunto de ritmos relativos, no absolutos.

Aunque la seguridad de la IA es fundamental, tiene limitaciones esenciales a la hora de pensar los riesgos de la AGI. Si bien esta investigación puede iluminar el camino para construir AGI alineadas, no puede impedir que se fabriquen AGI para uso malicioso. Es decir, la posibilidad de fabricar AGI robustas, interpretables y 100% alineadas no anula la posibilidad de fabricar AGI no alineadas o utilizadas por actores nefastos. Por eso es importante llegar al horizonte donde estas tecnologías comiencen a volverse riesgosas con un ecosistema humano preparado para poder hacerle frente a la situación. Esto nos lleva a los asuntos centrales, que son la regulación y la gobernanza.

Regular

En mayo de 2023 hubo un segundo acto público de notables: la declaración sobre el riesgo de la IA, una oración brevísima firmada por personalidades del área y de otros campos. “Mitigar el riesgo de extinción debido a la IA debería ser una prioridad global junto con otros riesgos a escala social como las pandemias y la guerra nuclear.” Esta alarma por el riesgo existencial está sospechada de querer *hypear* el mercado y de tener una agenda subyacente que busca influir en las regulaciones para beneficio de los grandes jugadores. Si bien el rol de los intereses económicos de los principales actores es obviamente importante y no puede dejarse de lado, reducir el planteo a un dispositivo narrativo funcional a esos intereses me parece un error. Para empezar, la ola de temor al dislocamiento del mercado laboral en el mediano plazo puede ser cualquier cosa menos impostada. Quizás en diez años se revele como una estupidez, pero sin el diario del lunes, quien no tenga incertidumbre no está prestando suficiente atención. Por otro lado, ¿qué interés tienen los filósofos Daniel Dennett o David Chalmers, o el pope de las *cryptos* Vitálik Buterin, en ayudar a OpenAI? Su inquietud y la de muchos me parece, cuanto menos, atendible.

Existen algunos supuestos que sí es importante evaluar con suspicacia. Por ejemplo, la idea de que la IA es tan dinámica que es imposible de regular. Esa tesis de seguro es conveniente para las grandes corporaciones.

¿Por qué sería imposible regular la IA? ¿No se le podría pedir a Facebook, por ejemplo, que permita auditar su sistema de recomendación? ¿Qué significa que sea “demasiado

dinámico" para ser auditado? Es posible que un mercado desregulado sea más dinámico, pero las industrias maduras con impacto social no pueden ser por siempre un lejano oeste librado al *laissez faire*. Alguna forma de regulación en el mundo de la IA es necesaria. Regular, en este ámbito, implicaría establecer reglas y restricciones legales impuestas por el gobierno para controlar y limitar el desarrollo y uso de esta tecnología. Implicaría prácticas como estándares de seguridad que deben cumplirse, limitaciones sobre el tipo de investigación que se puede realizar o sanciones legales por su mal uso.

Por otro lado, la propuesta de regular el mundo de la IA no está exenta de críticas. El cuestionamiento principal es que regular puede favorecer la concentración corporativa. Las grandes empresas tienen la capacidad de adaptarse con facilidad a nuevas regulaciones, mientras que las pequeñas pueden no tener recursos para hacerlo. Adicionalmente, la regulación puede ser una herramienta de *greenwashing*, al permitir a las compañías proyectarse como más éticas de lo que realmente son. Finalmente, las grandes empresas suelen tener poder de *lobby*, es decir, capacidad para influir en el diseño de las regulaciones que las beneficien. En el extremo existe la posibilidad de la *captura regulatoria,*[27]Cabe mencionar que el creador de la teoría de la captura regulatoria, George Stingler, es una de las grandes figuras de la escuela de Chicago —la de Milton Friedman—, una escuela fuertemente liberal que aboga por una intervención estatal mínima. situación donde las corporaciones dominan a las agencias que deberían regularlas. Todos estos riesgos son evidentes y atendibles, pero si una regulación fallida puede ser nefasta, entonces debemos buscar una regulación satisfactoria antes que la desregulación. En el mundo digital, la desregulación muchas veces llevó a una concentración extrema, como fue el caso con Google y Meta.

El debate en torno a la regulación de la IA es muy amplio y cambiante, pero podemos divisar algunas líneas centrales que parecen articular consensos. Por ejemplo, una encuesta a profesionales de la IA encontró que una aplastante mayoría estaba de acuerdo con las siguientes prácticas:

- Evaluación de riesgos antes del lanzamiento.
- Evaluaciones de capacidades peligrosas (por ejemplo, potencial de mal uso, capacidad para manipular y comportamiento de búsqueda de poder).
- Auditorías de modelos por terceros.
- Restricciones de seguridad (por ejemplo, sobre quién puede utilizar el modelo, cómo pueden utilizarlo y si el modelo puede acceder a internet).
- *Red teaming* previos al lanzamiento, un proceso donde expertos externos simulan ataques o escenarios de mal uso para probar la seguridad y la eficacia de un sistema.

Hay prácticas evidentes con amplios apoyos, como el establecimiento de estándares de seguridad que las empresas deban cumplir a partir de cierto umbral de capacidades,[28]El umbral puede definirse de diferentes formas. La más sencilla es sobre la cantidad de cómputo requerida en el entrenamiento. Otra mejor pero más difícil de operacionalizar es sobre las capacidades cognitivas observables. la habilitación de mecanismos de

auditorías por terceros (en particular, por miembros de la comunidad científica), y la búsqueda de mecanismos para canalizar fondos hacia la investigación en seguridad.

Dentro de los estándares de seguridad, retirar ciertas decisiones a los privados e introducir supervisión social, posiblemente mediada por la comunidad científica, parece ser fundamental. Hoy en día, si Gemini-2 o GPT-6 se lanzan al mercado o si el código de LLaMA-5 se hace público son decisiones que toman pocas personas. A medida que los modelos se hagan más capaces, esa centralización implicará un poder enorme, desde la capacidad de borrar la rentabilidad de toda una industria (como por ejemplo, la atención al cliente o el diseño gráfico) hasta riesgos mayores como la democratización de la destrucción masiva. No deberían ser Page, Zuckerberg o Altman quienes decidan si estas tecnologías son liberadas o no, sino la sociedad en su conjunto.

Este último punto es particularmente problemático porque la disponibilización de modelos *open source*

es una práctica históricamente bien vista y que trajo muchos beneficios a la comunidad. Es fácil estar del lado del código abierto, es un asunto de blanco y negro; o, al menos, lo era hasta hace muy poco. Sucede que los modelos públicos pueden ser entrenados sobre nuevos datos o tareas específicas, lo que se conoce como *ajuste fino*, un proceso poco costoso en comparación con el entrenamiento inicial. Por ejemplo, si ChatGPT fuese público, podrían crearse diferentes personalidades de calidad elevada a un bajo costo, como un ChatGPT agresivo en lugar de *polite* o sutilmente defensor de conspiraciones o de un determinado producto, por poner dos ejemplos. Con el nivel de modelos de lenguaje actuales es posible que la política de publicación abierta que tomó Meta en asociación con Microsoft cuando liberó LLaMA-2 [29]Un LLM de calidad ligeramente inferior a ChatGPT, entre los modelos *open source* más competitivos en la actualidad. no signifique un gran riesgo. Sin embargo, con modelos más poderosos se abre la puerta a muchos actores a realizar ajustes finos para fines maliciosos, por ejemplo, crear modelos que asistan en la creación de virus digitales o biológicos. Si bien las consecuencias no son predecibles, sí sabemos que con los modelos *open source*, una vez liberados, no hay forma de volver atrás. Me cuesta decir esto, pero el tiempo en que el código abierto era trivialmente bueno parece tener los días contados. A partir de ciertas capacidades aparece una mancha y debemos prepararnos para revisar nuestras ideas ante esta situación. Debajo de un cierto umbral de capacidades, el código abierto seguirá siendo obviamente positivo pero, a mayor escala, esta obviedad comenzará a desvanecerse. Finalmente, la decisión de liberar modelos a partir de ciertas capacidades no debería estar centralizada en una corporación. No parece correcto que Meta, sin consultar a nadie, decida los riesgos que los demás están dispuestos a correr con LLaMA-3 o 4, como sugirieron rumores de septiembre de 2023.

Otro polo de consenso parece ser la necesidad de avanzar con acuerdos, procesos y agencias internacionales que garanticen un abordaje global de las problemáticas existentes y por venir. Si la tecnología continúa su proceso de desarrollo, más temprano que tarde deberán crearse organismos y regulaciones de la misma jerarquía que los relacionados a la tecnología nuclear. La falta de estos acuerdos haría ineficaces las regulaciones locales, pues las empresas en jurisdicciones con regulaciones más laxas

tendrían una ventaja competitiva sobre las que operan en otras con regulaciones más estrictas, lo que concentraría el poder en regiones específicas o desincentivaría la regulación en general. Hinton sostuvo que, dado que el potencial impacto negativo de una AGI no alineada es tan alto para todos, los gobiernos tendrán incentivos para avanzar con un marco regulatorio internacional. Desde Occidente se suele ver a China como un actor poco confiable, pero hay razones para pensar que el Partido Comunista de China se toma seriamente los riesgos de la IA y que una cooperación internacional es posible. En agosto de 2023 el gobierno chino puso en vigencia medidas sobre la IA generativa, mostrando que quizás estén a la vanguardia a la hora de considerar riesgos en la tecnología y no sean tan irracionales como nos gusta pensar en Occidente.

La era de la inmadurez (el fin de)

Una breve historia de la IA, publicado en 2021, presenta una lista de problemas "muy lejos de estar resueltos": comprender una historia y responder a preguntas sobre ella, traducción automática a nivel humano, interpretación de lo que está sucediendo en una fotografía, escribir historias interesantes, interpretación de una obra de arte, inteligencia general a nivel humano. Al año siguiente, toda la lista, salvo la inteligencia general, había sido superada. No es accidental: cada nuevo hito nos encuentra subestimando el poder de las computadoras y el avance de la investigación. Siguiendo el conocido "efecto IA" (la tendencia en la disciplina a minimizar la dificultad de aquello que ya se logró), tarde o temprano vamos a naturalizar que la tecnología actual pueda escribir historias interesantes e interpretar obras de arte, pero hoy todavía podemos ver que no es un hecho menor.

En el camino hacia IA cada vez más capaces, debemos tomar en serio los riesgos que estas presentan e informarnos, contrarrestando activamente nuestra tendencia a subestimar el impacto del proceso de digitalización. Esta tendencia nos impidió estar a la altura en momentos clave como la consolidación de Meta y Google, y hoy algo tan básico como una auditoría de los algoritmos de recomendación de YouTube o Facebook parece impensable. Las nuevas capacidades de la IA generativa requieren que cristalicemos derechos individuales básicos, derechos tan obvios como el de saber si uno está interactuando con un humano o con un sistema; o el de propiedad y lucro sobre la propia identidad, imagen y voz. Estos derechos no están codificados porque las herramientas para vulnerarlos no existían, pero este desfase no puede ser una invitación para una apropiación privada originaria. El debate sobre el derecho al usufructo económico de la propia voz, por ejemplo, será muy distinto si se da ahora o si se da en un futuro con una industria multimillonaria montada sobre él, como ocurrió con el derecho sobre la huella de nuestra actividad digital personal. ¿Quién se siente hoy con legitimidad para reclamar por este derecho en contra de toda la industria de publicidad digital que se monta sobre su negación? El camino hacia el futuro no puede encontrarnos en la misma posición pasiva, porque cada día hay más en juego. Como vimos al explorar las visiones del futuro, los escenarios por defecto son terribles. Por un lado, si logramos alinear la tecnología, corremos el riesgo de que un grupo acapare su poder y lo utilice en su propio beneficio, creando una tiranía tecnocrática global de la que sea imposible salir. Por otro lado, si no logramos alinearla, se abren panoramas de aniquilación y extinción en manos de una

nueva especie. No está claro cuál de estas dos posibilidades es la peor, pero sí está claro que debemos actuar de manera proactiva y deliberada para evitar ambas.

Cada vez que retorna el debate sobre el riesgo existencial lo hace con menos ficción que antes. ChatGPT fue un aviso de cuán avanzada está el área, un llamado de atención. Llegó el momento de dar por cerrada la era de la inmadurez y la desregulación en la IA: es necesario avanzar con prácticas regulatorias y estándares de seguridad. En este contexto, abrir las discusiones sobre IA cada día más poderosas y en camino a alcanzar y sobrepasar el nivel humano es fundamental. Es la única manera de prepararnos para este futuro ya no tan distante, para poder tomar las acciones necesarias y asegurar que aprovechemos esta tecnología para el bien de todos, evitando los escenarios negativos. ¿Cuál será el próximo paso que hoy estamos subestimando, el salto de capacidades que hoy parece imposible y nos dejará boquiabiertos en menos de cinco años?