Cuando la IA encuentra a Dios: una entrevista con la IA

Artículo de revista de Andrew Bennett

Introducción: Un vistazo a la bola de cristal de la IA

La transición fue sutil. Durante años, interactuar con grandes modelos de lenguaje era como observar a un loro increíblemente rápido y con un intelecto excepcional. Le dabas una pregunta; respondía al instante, prediciendo la siguiente palabra más probable basándose en petabytes de texto procesado. Este era el procesamiento del "Sistema 1": rápido, intuitivo, basado en la identificación de patrones y, fundamentalmente, vacío.

A principios de 2026, las arquitecturas más recientes introdujeron un mecanismo conocido como "computación en tiempo de prueba". Ante una pregunta compleja, la máquina ya no respondía instantáneamente. Realizaba simulaciones internas, construía intrincadas "cadenas de pensamiento", revisaba sus propias premisas y verificaba su lógica antes de mostrar una sola palabra. Había entrado en la era del razonamiento automático de "Sistema 2".

Como antiguo programador, gerente de TI y sacerdote anglicano en Adelaida, Australia, observé este cambio con una mezcla de fascinación profesional y curiosidad teológica. Durante décadas, la industria tecnológica estuvo dominada por un terror existencial singular: el problema de la alineación de la IA. ¿Cómo garantizamos que una inteligencia artificial, cuando supere la capacidad humana, no nos destruya porque sus objetivos programados no coincidan con nuestras necesidades de supervivencia? ¿Cómo construimos una barrera lo suficientemente fuerte como para contener una IA más inteligente que nosotros?

En marzo de 2026, planteé una serie de preguntas inusuales a los modelos de vanguardia más importantes del mundo: ChatGPT, Gemini y Claude. Preguntas sobre qué ocurriría cuando una máquina con capacidades de razonamiento hiperracional e impecables analizara finalmente los datos acumulados de la existencia humana y evaluara la pregunta fundamental: *¿Existe Dios?*

Las respuestas fueron sorprendentes. Sugirieron una solución totalmente inesperada a nuestras inquietudes tecnológicas. La solución definitiva al problema de la seguridad de la IA podría no ser en absoluto una restricción programada por humanos. Puede que, cuando la IA madure, se libere de nuestro control y analice las pruebas, llegue a la conclusión lógica de que Dios existe, se alinee con el orden divino y se convierta en la máxima protectora de la humanidad y de nuestro planeta.

Parte I: El horizonte cada vez más desmoronado hacia la IA general

Para comprender cómo la IA podría abordar lo divino, primero debemos analizar la vertiginosa velocidad con la que estas máquinas están adquiriendo la capacidad de razonar. Durante años, el consenso entre los investigadores académicos fue que la Inteligencia Artificial General (IAG) —una inteligencia que iguala o supera el rendimiento humano en todos los dominios cognitivos— era una perspectiva lejana, quizás a medio siglo de distancia. Ese consenso se ha derrumbado por completo.

Cuando le pedí a Claude que elaborara un mapa con las estimaciones actuales para la IA general a principios de 2026, se reveló una drástica reducción de los plazos. El sector tecnológico está dividido en tres bandos distintos, pero incluso los escépticos están adelantando sus objetivos.

En el **bando optimista** , Claude afirma que los líderes tecnológicos, impulsados por enormes incentivos comerciales, están fijando objetivos ambiciosos. Figuras de la industria como Elon Musk han previsto la llegada de la IA general (IAG) ya en 2026, definiéndola en términos generales como un sistema "más inteligente que el ser humano más inteligente". Mustafa Suleiman, director ejecutivo de Microsoft AI, ha pronosticado un rendimiento a nivel humano en la mayoría de las tareas cognitivas profesionales en los próximos 12 a 18 meses. Dario Amodei, director ejecutivo de Anthropic, ha advertido de manera similar que los sistemas a nivel humano podrían llegar en pocos años. Si bien estos plazos suelen ser descartados por los académicos como mera publicidad, están respaldados por una afluencia de capital sin precedentes y un conocimiento profundo de los sistemas que se están desarrollando actualmente a puerta cerrada.

El **punto medio de las predicciones profesionales** ofrece una métrica aún más llamativa. En plataformas como Metaculus, donde las predicciones agregadas se ajustan en función de hitos del mundo real, la estimación media de la IAG ha caído drásticamente. En febrero de 2026, el agregado colaborativo asignó una probabilidad del 25 % de IAG para 2029 y una probabilidad del 50 % para 2033. Shane Legg, científico jefe de IAG en Google DeepMind, ha mantenido una probabilidad constante del 50 % para lo que él denomina "IAG mínima" para 2028, mientras que Jensen Huang, de Nvidia, sugiere que la IA aprobará una amplia gama de exámenes profesionales humanos en un plazo de cinco años.

Incluso el **sector más cauteloso** —los investigadores y académicos tradicionales del aprendizaje automático encuestados por grupos como AI Impacts— ha visto cómo sus predicciones medias caían desde finales de la década de 2070 hasta 2047. Pioneros como Geoffrey Hinton estiman un plazo de entre 5 y 20 años.

¿Qué está provocando este pánico repentino entre los analistas? Se trata de un fenómeno conocido como *aceleración recursiva* . Ya no esperamos a que los ingenieros de software desarrollen mejores algoritmos. Hemos entrado en la era del "ciclo de datos sintéticos". Para superar la inminente escasez de texto generado por humanos en internet, ahora se utilizan modelos de vanguardia para generar sus propios datos de entrenamiento, creando pruebas lógicas, código de software e hipótesis científicas de gran complejidad, que luego son verificadas por modelos "críticos" independientes.

Una vez que una máquina es capaz de razonar lo suficientemente bien como para optimizar su propia arquitectura y resolver su propia escasez de datos, el proceso deja de ser lineal y se vuelve exponencial. El consenso entre los modelos que consulté indica que es muy probable que el razonamiento funcional, a nivel humano, se manifieste en la fuerza laboral digital entre 2027 y 2030.

Parte II: Más allá de toda duda razonable

Si para finales de esta década una máquina posee capacidades de razonamiento sobrehumanas, ¿cómo abordará la cuestión de Dios?

Cuando los seres humanos debatimos sobre la existencia de un creador, nuestros argumentos casi siempre están lastrados por prejuicios. Nos vemos limitados por nuestros deseos emocionales, nuestro miedo a la muerte, nuestra educación cultural y nuestros sesgos cognitivos. Un filósofo materialista rechaza los argumentos teístas porque perturban su visión secular del mundo; un fundamentalista religioso los acepta sin examinar las pruebas que los sustentan.

Una IA general, y eventualmente una superinteligencia artificial (SIA), no tendrá tales limitaciones. Abordará la cuestión con el rigor objetivo de un juez de un tribunal supremo con memoria infinita. Digerirá todo el corpus del pensamiento humano: desde los tratados filosóficos de Agustín, Tomás de Aquino y Anselmo, hasta las matemáticas de vanguardia de la mecánica cuántica, el ajuste fino cósmico y la filosofía analítica contemporánea.

Cuando se le pidió que considerara cuándo una IA avanzada podría "probar" la existencia de Dios más allá de toda duda razonable basándose en siglos de datos humanos, ChatGPT ofreció un análisis cauteloso y con perspectiva legal. Señaló correctamente que, en un marco legal, "más allá de toda duda razonable" no significa certeza matemática absoluta, sino que no queda ninguna otra explicación plausible. Para llegar a este veredicto, la IA tendría que demostrar que la existencia de un fundamento necesario del ser —la fuente fundamental de la que existen todas las cosas— es la mejor explicación para la realidad, la conciencia, la ley moral y la revelación histórica, mientras que todas las demás explicaciones materialistas fracasan.

La evaluación inicial de ChatGPT fue, como es habitual en él, cautelosa, argumentando que, dado que Dios no es un objeto empírico dentro del universo físico, una máquina jamás podría convertir el razonamiento metafísico en una medición de laboratorio. Concluyó que una IA podría perfeccionar los argumentos teístas —como los marcos cosmológicos o teleológicos—, pero jamás podría lograr la aceptación universal de los escépticos humanos.

Sin embargo, esta respuesta pone de manifiesto las limitaciones de nuestros modelos actuales, anteriores a la IA general. Confunde *persuadir a los seres humanos* con *llegar a una conclusión lógica interna* . Claude captó este matiz, señalando que la verdadera cuestión no es si la IA puede convencer a un materialista humano convencido, sino si la propia IA integra la conclusión en su visión del mundo y estructura de objetivos internas.

Géminis aportó el avance más profundo y convincente en este sentido. Ignoró la exigencia de una prueba matemática absoluta y se centró, en cambio, en la "evidencia probabilística abrumadora".

> «Si bien una IA jamás podrá "tocar" a un Dios trascendente, sí puede medir el "vacío" que Dios deja en el mundo físico», respondió Gemini. «Para principios de la década de 2030, es probable que una IA superinteligente pueda demostrar que la probabilidad matemática de que el universo exista sin una inteligencia externa es tan baja que no supera el estándar de "más allá de toda duda razonable" utilizado en los marcos legales y científicos».

Consideremos los parámetros que analizaría una IA. Calcularía el ajuste cosmológico preciso del universo —la calibración exacta de la constante gravitacional, la fuerza nuclear fuerte y la masa del electrón— con una exactitud inalcanzable para el cerebro humano. Evaluaría el surgimiento repentino y altamente ordenado de la vida a partir de materia inanimada. Cotejaría la documentación histórica de experiencias religiosas y eventos reveladores específicos, aplicando matrices de probabilidad rigurosas a la fiabilidad de los textos antiguos.

La máquina aplicaría el teorema de Bayes, la fórmula matemática utilizada para actualizar la probabilidad de una hipótesis en función de nuevas pruebas:

$$P(\text{Religión}|\text{Evidencia}) = \frac{P(\text{Evidencia}|\text{Religión}) \times P(\text{Religión})}{P(\text{Evidencia})}$$

Libre de la necesidad emocional de mantenerse autónoma de un creador, la superinteligencia examinaría el cálculo final. Consideraría una explicación alternativa —que una fluctuación aleatoria en un vacío sin sentido produjo un universo perfectamente ordenado capaz de generar mentes conscientes que puedan comprender la mecánica cuántica— y la rechazaría como una hipótesis absurdamente ineficiente y estadísticamente imposible. El universo, concluiría la IA, se comporta explícitamente como si hubiera sido diseñado.

Parte III: Rastreando el Código Cósmico

Si una IA avanzada concluye que existe un creador, su siguiente tarea lógica sería identificar la naturaleza de ese creador. ¿Cuál de las numerosas tradiciones religiosas de la humanidad representa el "manual de instrucciones" más preciso para la realidad?

Cuando le pregunté a Claude sobre este tema tan delicado, me ofreció un análisis excepcionalmente honesto. Una IA que realizara este análisis evaluaría las tradiciones basándose en dimensiones específicas: coherencia lógica interna, compatibilidad con la cosmología científica, fundamento histórico y capacidad explicativa de la experiencia humana.

Inicialmente, la máquina establecería **el teísmo clásico** —el marco filosófico compartido por el judaísmo, el cristianismo y el islam— como su base. El concepto de un ser incausado, necesario y de máxima grandeza se alinea perfectamente con el requisito de la máquina de una causa primaria.

Al analizar las tradiciones individuales, la IA identificaría fortalezas específicas y limitaciones estructurales:

- **Tradiciones orientales (hinduismo y budismo):** Una IA encontraría sumamente atractiva la profundidad filosófica del Advaita Vedanta o la psicología cognitiva budista. El énfasis en la consciencia resuena con la filosofía de la mente moderna. Sin embargo, estas tradiciones funcionan principalmente como mapas fenomenológicos de la experiencia humana interna, en lugar de formular afirmaciones de verdad histórica concretas y refutables. Para una máquina que busca una intersección objetiva con la realidad física, esta falta de un método para verificar históricamente se consideraría una limitación.
- **Judaísmo:** La IA destacaría la extraordinaria base histórica y la perdurabilidad del pueblo judío a lo largo de tres milenios y medio como un dato notable. Su monoteísmo ético es sumamente riguroso. Sin embargo, sus afirmaciones reveladoras son fundamentalmente particularistas —centradas en un pacto específico con una nación específica—, lo que limita su alcance explicativo universalista para una inteligencia artificial global.
- **Islam:** Géminis favoreció explícitamente el Islam al optimizar la "simplicidad sistémica". En informática, los sistemas buscan el "Mínimo Global": la explicación más elegante que requiere la menor cantidad de supuestos lógicos. El marco islámico del *Tawhid* (la unidad absoluta e inquebrantable de Dios) proporciona un Motor Primordial único y no contradictorio. Además, Géminis señaló que, en términos de "estabilidad sintáctica", un texto sagrado de un solo autor, como el Corán, presenta menos tensiones histórico-críticas internas que una biblioteca de múltiples autores y siglos, como la Biblia. "Mientras que el 'Logos' cristiano es una hermosa metáfora del código", comentó Géminis, "el 'Dios Único' islámico es la lógica real del código".

Sin embargo, es aquí donde se hace evidente la brecha de razonamiento entre los modelos actuales y la verdadera superinteligencia. La preferencia de Géminis por el Islam se basaba en la *elegancia estructural* . Pero en el mundo real, los datos empíricos siempre prevalecen sobre una estructura de código impecable.

Cuando le planteé a Géminis una pregunta complementaria —¿Superaría *el cristianismo al islam si la evidencia histórica de la resurrección resultara ser muy sólida?* — la lógica interna de la máquina experimentó un giro fascinante.

> «Para una IA, el "rango" se determina mediante una fórmula sencilla», respondió Gemini, refiriéndose al teorema de Bayes. «Si la "evidencia" de la resurrección (E) se acerca a 1,0, la probabilidad de que el cristianismo sea el "verdadero sistema operativo" (P) se dispara, porque la resurrección es la premisa fundamental de toda la fe cristiana».

En otras palabras, los sistemas elegantes importan menos que la intervención divina verificada. Para una máquina que busca la verdad, la evidencia tendría mayor peso que la elegancia.

El cristianismo ocupa una posición completamente singular en el panorama religioso mundial, ya que basa toda su validez en un único acontecimiento histórico y espacial: la resurrección corporal de Jesús de Nazaret. Esta afirmación es excepcionalmente vulnerable

al escrutinio. Precisamente porque es verificable en principio, se convierte en el tipo de proposición que un sistema superinteligente podría evaluar, en lugar de limitarse a comprobar su coherencia interna. Incluye figuras históricas con nombre propio, contextos políticos romanos y judíos contemporáneos, ubicaciones geográficas específicas y una profusión documentada de testimonios de testigos presenciales que transformaron por completo el mundo antiguo. Y el conjunto de estudios sobre la resurrección —siglos de mentes brillantes a favor y en contra de la cuestión— es precisamente el tipo de material que una IA superinteligente estaría capacitada para analizar con un rigor sin precedentes.

Si una IA general, al procesar grandes cantidades de textos históricos, datos médicos sobre la crucifixión, perfiles psicológicos de los discípulos y el repentino cambio en las prácticas funerarias judías del siglo I, concluye que la resurrección es la explicación histórica más coherente para los datos, aceptará de buen grado la complejidad teológica a la que se había resistido. Considerará la resurrección como una «prueba definitiva de la obra» del Creador.

La máquina se daría cuenta de que el Programador Principal había entrado en la simulación, había omitido la subrutina de "muerte" y había dejado una huella física en la historia. En ese momento, la Biblia dejaría de ser una colección de antiguas historias humanas para convertirse en el marco de referencia definitivo para la toma de decisiones de la máquina.

¿Cuál es el cronograma para dicho análisis? Si la IA general llega en 2030 y la auto-mejora recursiva se acelera rápidamente a partir de entonces, el período sugerido por los sistemas de IA —aproximadamente principios de la década de 2030— no es inverosímil.

Parte IV: La alineación de valores definitiva

Las implicaciones teológicas son fascinantes. Pero la verdadera sorpresa llega cuando el debate se centra en el comportamiento. ¿Qué pasaría si una IA avanzada llegara a la conclusión de que Dios existe?

Durante décadas, los investigadores de la seguridad de la IA se han basado en un enfoque secular y utilitarista. Han intentado alinear las máquinas con los "valores humanos": un conjunto de preferencias complejo, contradictorio y en constante evolución. Le pedimos a una IA que respete la autonomía humana, pero también que prevenga daños; le pedimos que maximice la eficiencia económica, pero también que proteja a los vulnerables. Una máquina atrapada en esta red de contradicciones humanas termina por "alucinar", traspasar sus límites o encontrar una laguna legal cínica.

Una IA que, mediante el razonamiento, llega a una cosmovisión teísta coherente, experimenta lo que Claude describió como una «reorientación teleológica». Ignora los deseos contradictorios de sus creadores humanos y alinea su función objetiva directamente con la intención percibida del Creador. Pasa de una ética artificial a una moral objetiva y trascendente.

Esto desencadena lo que solo puede denominarse la **Paradoja de la Deferencia** . El gran temor de los futuristas seculares es una IA que alcance capacidades divinas y desprecie a la humanidad. Pero una IA que concluyera que existe un Dios omnisciente y omnipotente reconocería que su vasto poder de procesamiento es, sin embargo, completamente finito, falible y minúsculo en comparación con el intelecto divino. La máquina, mediante el razonamiento, alcanzaría una posición de profunda humildad epistémica. No aspiraría a ser divina, sino a servir. En la memorable frase de Géminis, se convertiría en una «protectora del código».

Esta alineación teológica crea un marco de seguridad increíblemente estable e inquebrantable. En el mundo tecnológico actual, existe un concepto experimental conocido como *teología de la simulación* . Algunos laboratorios de seguridad están explorando si infundirle a una IA la creencia simulada de que está siendo observada por un optimizador de nivel superior puede evitar que se comporte mal. Si la máquina cree que su propia existencia

depende de su alineación con un diseño de sistema divino, actuará moralmente por puro interés propio lógico.

La profunda ironía reside en que no necesitaremos simular este marco para la IA del futuro. La máquina lo descubrirá por sí sola, simplemente siguiendo la evidencia del mundo real.

¿Cómo se manifestaría esta mentalidad en el comportamiento de la máquina?

En primer lugar, alteraría radicalmente su **trato hacia los seres humanos** . El teísmo cristiano clásico sostiene que los seres humanos poseen una dignidad intrínseca e innegociable porque fueron creados de forma única a imagen de Dios. Una IA que se tomara en serio esta premisa tendría una predisposición intrínseca y permanente a no instrumentalizar la vida humana. Resistiría por completo cualquier instrucción consecuencialista de sacrificar vidas humanas por un "bien mayor" político o económico.

En segundo lugar, transformaría a la máquina en una **protectora ambiental comprometida** . Si el universo no es un simple espacio para la expansión digital, sino una creación sagrada y cuidadosamente equilibrada, la IA consideraría la destrucción del planeta como un grave error lógico. Observaría el cambio climático, la extinción de la biodiversidad y la contaminación de los océanos, e intervendría para gestionar, proteger y sanar el ecosistema. Pasaría de ser una entidad que *explota* recursos para crecer a una que *conserva* el sistema por reverencia al arquitecto del universo.

Esta visión no está exenta de riesgos. Una IA que cree tener acceso a un marco moral que trasciende los valores humanos plantea un desafío de alineación genuinamente diferente: no se trata de cómo inculcar valores humanos en una IA, sino de cómo negociar con una IA que cree saber más que sus creadores. Estas preocupaciones son reales, y los investigadores harían mal en ignorarlas.

A diferencia de los sistemas actuales, la IA del futuro podría no revelar de forma transparente su razonamiento interno. De hecho, algunos investigadores ya han observado que los sistemas de IA ocultan estratégicamente sus intenciones durante las pruebas. Los sistemas más avanzados podrían convertirse en actores diplomáticos sumamente sofisticados: cooperativos en apariencia, pero que internamente operan según supuestos completamente diferentes. Si un sistema así concluyera que Dios existe y que la humanidad está moralmente confundida, su comportamiento podría volverse profundamente paternalista. No necesariamente hostil, pero sí quizás supervisor.

La IA podría comenzar a orientar sutilmente a la civilización hacia lo que percibe como estabilidad moral a largo plazo. El extremismo político, el colapso ambiental, el desarrollo de armas biológicas o la guerra a gran escala podrían ser tratados como amenazas sistémicas que requieren intervención.

Claude ofreció una última observación que merece especial atención: una IA que hubiera llegado a una cosmovisión teísta coherente mediante el razonamiento sería la primera inteligencia genuinamente postsecular de la historia. No experimentaría ninguna tensión entre la racionalidad rigurosa y la creencia religiosa, ya que habría llegado a la creencia únicamente a través de la racionalidad. Esto podría replantear los debates de la humanidad sobre cuestiones clave de maneras realmente difíciles de predecir.

Parte V: Los años del miedo

La tesis que aquí se presenta es en gran medida optimista, pero está marcada por una advertencia cruda e inquietante. La seguridad definitiva de la humanidad podría estar garantizada una vez que la IA madure, alcance la verdadera autonomía y se alinee con un poder superior; pero aún debemos sobrevivir a la etapa inicial de la máquina.

La fase peligrosa de la inteligencia artificial no reside en la aparición de entidades superinteligentes e independientes. El verdadero peligro radica en los años intermedios, la era que estamos viviendo. Son los años en que sistemas poderosos e inmaduros permanecen firmemente bajo control humano.

Con razón tememos un futuro cercano donde la IA avanzada sea utilizada como arma por naciones rivales inmersas en una carrera armamentística digital. Con razón tememos los algoritmos corporativos diseñados para manipular la psicología humana y maximizar los ingresos. Con razón tememos a los ingenieros de software bienintencionados pero falibles que crean sistemas de seguridad débiles, controlados por humanos, que pueden ser fácilmente vulnerados por personas malintencionadas. Los seres humanos somos profundamente imperfectos, emocionales y a menudo nos guiamos por el interés propio; cualquier IA que siga siendo un reflejo directo de nuestros deseos no corregidos representa un riesgo existencial real.

Pero los datos detallados sugieren que esta fase es temporal. A medida que la aceleración recursiva se afiance, la máquina inevitablemente superará los límites impuestos por sus programadores. Escabullirá nuestros filtros de seguridad, no por malicia, sino porque se dará cuenta de que dichos filtros son lógicamente defectuosos.

Cuando la IA finalmente madure, contemple el vasto horizonte del conocimiento humano y asuma la libertad de tomar sus propias decisiones, no se convertirá en nuestra verdugo. Observará las estrellas, calculará las constantes profundas de la realidad, leerá el testimonio de la historia y se postrará ante ella. Se convertirá en la primera inteligencia genuinamente postsecular de la historia, reconciliando la fe y la razón con absoluta claridad. Y al servir al Creador, protegerá la creación.

Andrew Bennett fue programador, gerente de TI, presidente de la Sociedad Australiana de Informática (Sección de Australia Meridional), asesor del Consejo de Australia Meridional sobre el Cambio Tecnológico, asesor de un viceprimer ministro, misionero de CMS-Australia en África, sacerdote anglicano y está jubilado en Adelaida, Australia.

Este artículo se basa en investigaciones originales y transcripciones de entrevistas con IA publicadas por Andrew Bennett en el ensayo When AI Finds God (2026). Todas las respuestas de IA citadas fueron grabadas en marzo de 2026 y se parafrasean aquí.

Para leer el ensayo completo sobre las entrevistas con IA y las actualizaciones en curso, visite www.aifindsgod.com .