Rev Argent Cir. 2026;118(1):e-1964- http://dx.doi.org/10.25132/raac.v118.n1.1964

In ipso vivimus et movemur: inteligencia espacial, modelos del mundo y cirugía ultracompleja de élite como *benchmark* de límite superior para la AGI (Artificial General Intelligence) encarnada

In Ipso Vivimus et Movemur*: spatial intelligence, world models, and elite ultra-complex surgery as the upper limit benchmark for embodied AGI (Artificial General Intelligence)*

Enrique Díaz Cantón

Departamento de Oncología Médica e Inteligencia Artificial, Facultad de Medicina, Instituto Universitario CEMIC, Buenos Aires, Argentina

El autor declara no tener conflictos de interés.
Conflicts of interest None declared.

Declaración de uso de IA: el autor utilizó Claude 4.5 Opus (Anthropic, 2025) para asistir con la sintaxis del inglés y la edición del lenguaje. El autor mantiene plena responsabilidad por el contenido, precisión e integridad del trabajo.

Correspondencia *Correspondence:* Enrique Díaz Cantón. E-mail: ediazcanton@iuc.edu.ar

RESUMEN

Los modelos de lenguaje extensos han logrado una fluidez notable en el razonamiento simbólico y el diálogo, pero permanecen fundamentalmente limitados en la competencia sensoriomotora rica en contacto que los cirujanos desarrollan a través de décadas de práctica deliberada. Basándome en la teoría de la cognición encarnada y las propuestas contemporáneas de modelos predictivos del mundo, argumento que el obstáculo más significativo que resta hacia la inteligencia artificial general (AGI, por sus siglas en inglés *Artificial General Intelligence*) no es lingüístico sino físico: la capacidad de anclar la percepción, la planificación y la acción dentro de las restricciones implacables de la dinámica del mundo real. Como *benchmark* (parámetro de referencia estándar) ilustrativo de límite superior, propongo el "*Benchmark* de AGI Quirúrgica", el punto hipotético en el cual un sistema robótico autónomo podría realizar de manera segura y fiable procedimientos quirúrgicos ultracomplejos (trasplante multivisceral, resecciones hepatopancreatobiliares complejas, reconstrucciones microquirúrgicas asistidas por robot) igualando o superando a los cirujanos humanos de élite. Esta perspectiva descompone el *benchmark* en una escalera práctica de hitos y describe los requisitos técnicos –censado táctil, control complaciente, planificación jerárquica y modelos predictivos del mundo– necesarios para tender el puente entre la simulación y la realidad clínica.

Palabras clave: *inteligencia artificial encarnada, inteligencia espacial, modelos del mundo, arquitectura predictiva de embedding conjunto, transferencia de simulación a realidad, robótica quirúrgica, cirugía autónoma, aprendizaje sensoriomotor, control motor jerárquico.*

ABSTRACT

Although large language models have achieved remarkable fluency in symbolic reasoning and dialogue, they remain fundamentally limited in their sensorimotor competence, which surgeons develop through decades of deliberate practice. Based on the theory of embodied cognition and contemporary proposals for predictive models of the world, I argue that the most significant obstacle standing in the way of artificial general intelligence (AGI) is not linguistic, but physical: the ability to anchor perception, planning, and action within the relentless constraints of real-world dynamics. As an illustrative benchmark for the upper limit, I propose the "AGI benchmark in surgery"—the hypothetical point at which an autonomous robotic system could safely and reliably perform ultra-complex surgical procedures, such as multi-visceral transplantation, complex hepatopancreatobiliary resections, and robot-assisted microsurgical reconstructions, while matching or surpassing the abilities of elite human surgeons. This perspective breaks down the benchmark into a practical ladder of milestones and describes the technical requirements—tactile sensing, compliant control, hierarchical planning, and predictive models of the world—needed to bridge the gap between simulation and clinical reality.

Keywords: *embodied artificial intelligence; spatial intelligence; models of the world; predictive joint embedding architecture; simulation-to-reality transfer; robotic surgery; autonomous surgery; sensorimotor learning; hierarchical motor control.*

Recibido | *Received* 20-01-26
Aceptado | *Accepted* 17-03-26

ID ORCID: Enrique Díaz Cantón, 0009-0004-0967-448X.

"Porque en él vivimos, nos movemos y existimos". (Hechos 17:28).

Introducción

Los sistemas de inteligencia artificial contemporáneos escriben, traducen, resumen y asisten el razonamiento clínico con fluidez impresionante, pero tienen dificultades con las habilidades sensoriomotoras que los cirujanos expertos refinan durante años de entrenamiento: manipulación estable del tejido, coordinación precisa de instrumentos e interacción segura con la anatomía viva bajo presión de tiempo. Esta asimetría hace eco de la paradoja de Moravec: el razonamiento abstracto resulta computacionalmente tratable, mientras que la percepción y la acción diestra demandan recursos computacionales y experienciales extraordinarios[1]. A lo largo de este ensayo, empleo una heurística "80/20" –no como una constante medida sino como un dispositivo conceptual–para subrayar que gran parte de la experiencia humana es encarnada más que proposicional. El progreso hacia la inteligencia general, sostengo, requerirá agentes capaces de aprender modelos predictivos del mundo, planificar a través de múltiples horizontes temporales y cerrar el bucle sensoriomotor en dominios físicos consecuentes[2-4].

Por qué el lenguaje solo alcanza una meseta para la agencia encarnada

El lenguaje ofrece una interfaz poderosa hacia el conocimiento humano acumulado, pero no puede sustituir el bucle de retroalimentación continua que enlaza la percepción con la acción. Las tareas ricas en contacto demandan censado de alto ancho de banda (visión, propiocepción, fuerza, tacto), control rápido bajo incertidumbre y predicción precisa de cómo las intervenciones alteran el mundo físico. El marco de LeCun para la inteligencia de máquinas autónomas destaca los modelos predictivos del mundo y las arquitecturas de *embedding* (representación vectorial), conjunto que representa y anticipa transiciones de estado latente, permitiendo así una planificación que trasciende el mero emparejamiento de patrones[2]. En paralelo, las políticas generalistas de visión-lenguaje-acción aspiran a heredar conocimiento semántico del preentrenamiento a escala de Internet mientras adquieren control anclado de conjuntos de datos robóticos heterogéneos[5]. Estas direcciones de investigación convergentes sugieren que el cuello de botella crítico no es "más texto" sino un modelado del mundo más rico y una encarnación física más segura.

De la metáfora a la escalera de *benchmark*: qué significa realmente 'capaz de cirugía'

El *Benchmark* (parámetro de referencia estándar) de AGI Quirúrgica es deliberadamente extremo. La cirugía ultracompleja comprime en una sola arena casi todos los requisitos para la inteligencia encarnada: contacto continuo con tejido, dinámica fisiológica rápida, anticipación de variantes anatómicas, patología engañosa, tolerancias de precisión milimétricas y consecuencias catastróficas por error. Cualquier cirujano que haya enfrentado una lesión arterial inesperada durante un procedimiento de Whipple comprende visceralmente lo que significa "inteligencia encarnada bajo presión". Para hacer este *benchmark* accionable propongo una escalera de hitos susceptibles de evaluación bajo restricciones de seguridad controladas mucho antes de que la autonomía procedimental completa sea factible:

- ***Nivel 0 - Manejo básico de instrumentos:*** agarre, corte y sutura consistentes en modelos de banco inanimados con perfiles de fuerza reproducibles.
- ***Nivel 1 - Primitivas de manipulación tisular:*** identificación del plano de disección, maniobras hemostáticas y retracción en especímenes cadavéricos o *ex-vivo* que exhiben *compliance* realista.
- ***Nivel 2 - Pasos procedimentales guionizados:*** ejecución de anastomosis estandarizadas, linfadenectomía sistemática y márgenes de resección predefinidos dentro de envolventes de seguridad.
- ***Nivel 3 - Procedimientos adaptativos bajo supervisión:*** respuesta en tiempo real a hallazgos intraoperatorios (hemorragia, anatomía no anticipada) con supervisión humana y autoridad de veto.
- ***Nivel 4 - Autonomía procedimental completa para casos estándar:*** colecistectomía, apendicectomía, reparación de hernia inguinal completadas con intervención humana mínima.
- ***Nivel 5 - Resecciones oncológicas complejas:*** pancreatoduodenectomía, escisión mesorrectal total, nefrectomía radical con reconstrucción vascular, procedimientos que requieren integración de juicio y técnica.
- ***Nivel 6* - Benchmark *quirúrgico de élite:*** trasplante multivisceral, resección hepática ex-vivo con autotrasplante, reconstrucción microquirúrgica robótica de colgajo libre, procedimientos que demandan la integración perfecta de múltiples competencias subespecializadas al más alto nivel.

Requisitos técnicos para la inteligencia encarnada rica en contacto

Lograr incluso hitos intermedios demanda integrar percepción, control y aprendizaje dentro de un sistema unificado que mantenga la seguridad bajo incertidumbre irreducible.

Censado y encarnación

El censado táctil distribuido a través de las puntas de los instrumentos, transductores de fuerza-torque en las articulaciones, retroalimentación propioceptiva precisa y visión estereoscópica calibrada son prerrequisitos para inferir el estado del tejido, la complacencia mecánica y los límites anatómicos en tiempo real, capacidades que los cirujanos experimentados desarrollan a través de miles de casos.

Control complaciente y reflejos

El control de impedancia de baja latencia, la actuación mecánicamente complaciente y los comportamientos protectores de tipo reflejo deben operar continuamente para prevenir daño tisular mientras preservan la precisión durante la disección delicada y responden instantáneamente a hemorragia súbita –de manera similar a como las manos de un cirujano se estabilizan reflexivamente cuando se entra inadvertidamente en una arteria–.

Planificación jerárquica

Los cirujanos expertos operan simultáneamente en múltiples escalas temporales: reflejos de milisegundos, primitivas motoras de nivel de segundos y planificación estratégica de minutos a horas. Los agentes quirúrgicos autónomos probablemente requerirán arquitecturas de capas análogas –bucles de estabilización reactiva, módulos de habilidades de nivel medio (disección, anastomosis, hemostasia) y planificadores de alto nivel capaces de revisar los objetivos procedimentales a medida que se despliega la anatomía–.

Modelos predictivos del mundo

El aprendizaje de representación predictiva autosupervisado –ejemplificado por las arquitecturas predictivas de *embedding* conjunto como JEPA (por su sigla en inglés *Joint Embedding Predictive Architecture*)– ofrece un camino hacia la codificación de estados físicos latentes y el soporte de simulación interna sin la fragilidad de los modelos generativos *pixel-perfect*[2,3]. Tales modelos del mundo podrían permitir a un agente "ensayar mentalmente" las consecuencias de una maniobra quirúrgica antes de comprometerse irreversiblemente.

Transferencia de simulación a realidad

Los motores de física de alta fidelidad y la simulación a gran escala aceleran la iteración algorítmica, pero la cirugía auténtica involucra deformación de tejido blando, dinámica de sangrado y variación anatómica específica del paciente que ningún simulador captura fielmente aún. Tender este puente requerirá aleatorización de dominio agresiva, identificación cuidadosa del sistema e –inevitablemente–recolección de datos del mundo real estructurada bajo supervisión ética rigurosa[6].

Progreso reciente: hacia agentes generalistas y modelos del mundo

Varios sistemas recientes ofrecen bloques de construcción relevantes para esta agenda. Genie aprende de entornos virtuales interactivos a partir de video no etiquetado, proporcionando generación de mundos controlables por acción útil para entrenar agentes a través de dinámicas diversas[7]. SIMA escala el comportamiento de seguimiento de instrucciones a través de mundos simulados heterogéneos mediante una interfaz genérica, y SIMA 2 (por sus siglas en inglés, *Scalable Instructable Multiword Agent*) extiende esto con completar objetivos de múltiples pasos más complejos [8,9]. En robótica, π0 propone un modelo de flujo de visión-lenguaje-acción entrenado a través de múltiples plataformas de *hardware* para control de propósito general[5]. *WeatherNext* 2, aunque distante de la manipulación quirúrgica, demuestra que el modelado predictivo a gran escala de sistemas físicos complejos puede superar significativamente las líneas base tradicionales, una prueba de existencia de que los modelos de dinámica aprendidos tienen valor práctico[10]. Ninguno de estos sistemas se aproxima a la competencia quirúrgica, pero juntos esbozan una trayectoria de investigación plausible.

Línea temporal especulativa

Cualquier línea temporal para lograr el *Benchmark* de AGI Quirúrgica completo permanece profundamente especulativa, contingente en la maduración del *hardware*, avances en ingeniería de seguridad y marcos regulatorios en evolución. Un objetivo a corto plazo más tratable es el progreso sistemático a través de los Niveles 0-3 dentro de entornos controlados –simulación, laboratorios cadavéricos y quizá configuraciones de tejido vivo

altamente restringidas–. Los Niveles 4-6 demandarían avances transformadores en robustez táctil, actuación complaciente, inferencia anatómica en tiempo real y aprendizaje seguro bajo las consecuencias irreversibles del error quirúrgico. Ofrezco esta escalera no como profecía sino como una invitación a la evaluación incremental y medible.

Conclusión

La tesis de este ensayo es directa: la inteligencia que opera en el mundo físico está restringida por la física, el contacto y la consecuencia. La competencia lingüística constituye una dimensión importante de la inteligencia general, pero la competencia encarnada puede resultar el cuello de botella decisivo para la agencia autónoma. El *Benchmark* de AGI Quirúrgica funciona como un límite superior deliberadamente provocativo, un recordatorio de lo que los sistemas contemporáneos aún no pueden aproximar, y un camino estructurado de hitos intermedios por los cuales el progreso futuro podría medirse. El día en que un sistema autónomo complete de manera segura un trasplante multivisceral marcará no meramente un logro tecnológico sino una expansión profunda de lo que entendemos que es la inteligencia.

"Y el Verbo se hizo carne".
(Juan 1:14)

■ ENGLISH VERSION

"For in him we live and move and have our being."
(Acts 17:28)

Introduction

Although contemporary artificial intelligence systems can write, translate, summarize, and assist with clinical reasoning with impressive fluency, they struggle with the sensorimotor skills that expert surgeons refine over years of training. These skills include stable tissue manipulation, precise instrument coordination, and safe interaction with live anatomy under time pressure. This asymmetry reflects Moravec's paradox: abstract reasoning is computationally manageable, while perception and dexterous action require extraordinary computational and experiential resources[1]. Throughout this essay, I employ an "80/20" heuristic—not as a constant measure but as a conceptual device—to emphasize that much of human experience is embodied rather than propositional. Progress toward general intelligence, I argue, will require agents capable of learning predictive models of the world, planning across multiple time horizons, and closing the sensorimotor loop in consequential physical domains[2-4].

The reason why language only reaches a plateau for embodied agency.

Language offers a powerful interface to accumulated human knowledge, but it cannot replace the continuous feedback loop that links perception with action. Rich-contact tasks require high bandwidth sensing (vision, proprioception, strength and touch), rapid control under uncertainty, and accurate prediction of how interventions alter the physical world. LeCun's framework for autonomous machine intelligence highlights predictive world models and joint embedding architectures (vector representation) that represent and anticipate latent state transitions, enabling systems to anticipate outcomes and plan ahead, going beyond mere pattern matching[2]. At the same time, general vision-language-action policies seek to inherit internet-scale semantic knowledge built on top of a pre-trained model while acquiring anchored control of heterogeneous robotic datasets[5]. These converging lines of research suggest that the critical bottleneck is not "more text" but rather richer modeling of the world and more reliable physical embodiment.

From metaphor to the benchmark ladder: the true meaning of "surgical capability".

The AGI benchmark in surgery is deliberately extreme. Ultra-complex surgery combines almost all the requirements for embodied intelligence in a single arena: continuous contact with tissue, rapid physiological dynamics, anticipation of anatomical variations, misleading pathology, millimeter precision tolerances, and catastrophic consequences for errors. Any surgeon who has encountered an unexpected arterial injury during a Whipple procedure intimately understands the concept of "embodied intelligence under pressure." To make this benchmark actionable, I propose a ladder of milestones that can be evaluated under controlled safety constraints well before full procedural autonomy becomes feasible:

- ***Level 0 - Basic instrument handling:*** consistent grasping, cutting, and suturing using inanimate bench models with reproducible force profiles.
- ***Level 1 - Basic tissue manipulation techniques:***

identification of dissection planes, bleeding control, and retraction in cadavers or ex vivo specimens exhibiting realistic compliance.

- ***Level 2 - Scripted procedural steps:*** execution of standardized anastomoses, systematic lymph node dissection, and predefined clear resection margins.
- ***Level 3 - Supervised adaptive procedures:*** real-time response to intraoperative findings (bleeding, unanticipated anatomical variations) under human direct supervision and with the authority to veto actions.
- ***Level 4 - Full procedural autonomy for standard cases:*** cholecystectomy, appendectomy and inguinal hernia repair completed with minimal human intervention.
- ***Level 5 - Complex oncological resections:*** duodenopancreatectomy, total mesorectal excision, radical nephrectomy with vascular reconstruction-procedures requiring the integration of judgment and technique.
- ***Level 6 - Benchmark in elite surgery:*** multivisceral transplantation, ex vivo liver resection and autotransplantation, robot-assisted free-flap microsurgical reconstruction-procedures that require the seamless integration of multiple subspecialized skills at the highest level.

Technical requirements for rich-contact embodied intelligence

Achieving even intermediate milestones requires integrating perception, control, and learning within a unified system that maintains safety under irreducible uncertainty.

Sensing and embodiment

Distributed tactile sensing at the instrument tips, joint-level force–torque transduction, accurate proprioceptive feedback, and calibrated stereoscopic visualization are fundamental for real-time inference of tissue properties, mechanical compliance, and anatomical boundaries—competencies acquired by experienced surgeons over thousands of procedures.

Compliant control and reflexes

Low-latency impedance control, mechanically compliant performance, and reflexive protective behaviors must be continuously operational to prevent tissue damage while preserving precision during delicate dissection and responding instantly to sudden bleeding—similar to how a surgeon's hands reflexively stabilize when inadvertently entering an artery.

Hierarchical planning

Expert surgeons integrate behavior across multiple temporal scales, from millisecond reflexive control and second-level sensorimotor corrections to higher-order strategic planning unfolding over minutes to hours. Autonomous surgical agents will likely require analogous layered architectures-reactive stabilization loops, mid-level skill modules (dissection, anastomosis, hemostasis), and high-level planners capable of reviewing procedural goals as the anatomy unfolds.

Predictive models of the world

Self-supervised predictive representation learning -exemplified by joint embedding predictive architectures (JEPA)- offers a path toward encoding latent physical states and supporting internal simulation without the fragility of pixel-perfect generative models[2,3]. Such models of the world could allow an agent to "mentally rehearse" the consequences of a surgical maneuver before committing to it irreversibly.

Transferring simulation to reality

High-fidelity physics engines and large-scale simulation accelerate algorithmic iteration, but real surgery involves soft tissue deformation, bleeding dynamics, and patient-specific anatomical variation that no simulator can yet accurately capture. Bridging this gap will require aggressive domain randomization, careful system identification, and -inevitably- structured collection of real-world data under rigorous ethical supervision[6].

Recent progress: toward generalist agents and models of the world

Several recent systems offer relevant building blocks for this agenda. Genie learns interactive virtual environments from unlabeled videos, generating action-controllable virtual worlds useful for training agents through diverse dynamics[7]. SIMA (Scalable, Instructable, Multiworld Agent) scales instruction-following behavior across heterogeneous simulated worlds using a generic interface, and SIMA 2 extends this by completing more complex, multi-step objectives[8,9]. π0 proposes a vision-language-action flow model trained across multiple hardware platforms for general-purpose robot control[5]. While far from surgical manipulation, WeatherNext 2 demonstrates that large-scale predictive modeling of

complex physical systems can significantly outperform traditional baselines, evidence that learned dynamics models have practical value[10]. While none of these systems compares to surgical competence, they collectively outline a plausible path for future research.

Hypothetical timeline

Any timeline for achieving the full AGI benchmark in surgery is deeply hypothetical and contingent on hardware development, advances in safety engineering, and evolving regulatory frameworks. A more manageable short-term goal would be to make systematic progress through Levels 0-3 in controlled environments, such as simulations, cadaver laboratories, and possibly highly restricted living tissue configurations. Levels 4-6 would require transformative advances in tactile robustness, compliant performance, real-time anatomical inference, and safe learning under the irreversible consequences of surgical error. I offer this ladder not as a prophecy but as an invitation to incremental and measurable evaluation.

Conclusion

The thesis of this essay is straightforward: intelligence operating in the physical world is constrained by physics, contact, and consequence. While linguistic competence is a primary pillar of general intelligence, embodied competence may represent the ultimate bottleneck in the development of autonomous agency. The AGI benchmark in surgery serves as a provocative upper bound—a stark reminder of the limitations of contemporary systems and a rigorous framework of intermediate milestones to measure future progress. The day an autonomous system safely performs a multivisceral transplant will not only be a technological achievement, but also a profound expansion of our understanding of intelligence.

"And the Word became flesh."
(John 1:14)

Referencias bibliográficas /References

1. Moravec H. Mind children: the future of robot and human intelligence. Cambridge (MA): Harvard University Press; 1988.
2. LeCun Y. A path towards autonomous machine intelligence. OpenReview [Internet]. 2022 [citado 2025 Dic 23]. Disponible en: https://openreview.net/pdf?id=BZ5a1r-kVsf
3. Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat MG, et al. Self-supervised learning from images with a joint-embedding predictive architecture. En: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. pp.15619--29.
4. Brooks RA. Intelligence without representation. Artif Intell. 1991;47(1-3):139-59. doi:10.1016/0004-3702(91)90053-M.
5. Black K, Brown N, Driess D, Esmail A, Equi M, Finn C, et al. π0: a vision-language-action flow model for general robot control. arXiv [Preprint]. 2024 [citado 2025 Dic 23]. Disponible en: https://arxiv.org/abs/2410.24164
6. Todorov E, Erez T, Tassa Y. MuJoCo: a physics engine for model-based control. En: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2012. pp.5026-33.
7. Bruce J, Dennis M, Edwards A, Parker-Holder J, Shi Y, Hughes E, et al. Genie: generative interactive environments. arXiv [Preprint]. 2024 [citado 2025 Dic 23]. Disponible en: https://arxiv.org/abs/2402.15391
8. Abi Raad M, Ahuja A, Barros C, Besse F, Bolt A, Bolton A, et al; SIMA Team. Scaling instructable agents across many simulated worlds. arXiv [Preprint]. 2024 [citado 2025 Dic 23]. Disponible en: https://arxiv.org/abs/2404.10179
9. SIMA Team. SIMA 2: a generalist embodied agent for virtual worlds. arXiv [Preprint]. 2025 [citado 2025 Dic 23]. Disponible en: https://arxiv.org/html/2512.04797v1
10. Google. WeatherNext 2: our most advanced weather forecasting technology. Google Blog [Internet]. 2025 Nov 17 [citado 2025 Dic 23]. Disponible en: https://blog.google/technology/google-deepmind/weathernext-2/