Received: 9 May 2026 | Revised: 10 July 2026 | Accepted: 24 July 2026 | Published online: 17 August 2026

REVIEW

Journal of Climbing and Walking Robots
2026, Vol. 00(00) 1–10
DOI: 10.47852/bonviewJCWR620210350

Movemur: Large World Models, Newtonian Physics, and the Embodied Bottleneck of Artificial General Intelligence

Enrique Díaz Cantón[1,*]

[1] *Department of Medical Oncology and Artificial Intelligence, School of Medicine, CEMIC University Institute, Argentina*

Abstract: Contemporary large language models achieve remarkable competence in symbolic reasoning, dialog, and the synthesis of internet-scale text and yet remain conspicuously deficient in the contact-rich, gravity-bound competence that any biological six-month-old begins to acquire while attempting to climb out of a crib. Drawing on embodied cognition, on Yann LeCun's joint-embedding predictive architecture (JEPA) and its video extension V-JEPA 2, and on Google DeepMind's interactive world-model line (Genie, SIMA, SIMA 2), this review argues that the principal remaining bottleneck on the path to artificial general intelligence (AGI) is not linguistic or symbolic but Newtonian: the capacity to ground perception, planning, and action within the constraints of classical mechanics as they impinge on a physical body in real time. We propose that climbing and walking robots occupy a privileged position in this research program because locomotion compresses into a single integrated task virtually every requirement for embodied intelligence under physical law while imposing failure modes that are immediate and irreversible. Building on prior surgical and grappling benchmarks, we introduce a Pauline diagnostic triad (vivimus/movemur/sumus), review recent empirical landmarks of legged-robot learning, outline a milestone ladder for locomotion-grounded embodied AGI, and frame the field's stance toward the sim-to-real gap. The thesis is concise: until an autonomous system can climb an unfamiliar staircase, traverse rubble, and recover from a slip with adult competence, claims of general intelligence remain rhetorical. We advance this as a falsifiable hypothesis: walking and climbing are plausibly necessary, though not sufficient, conditions for it.

Keywords: embodied artificial intelligence, world models, joint-embedding predictive architecture, legged locomotion, climbing robots

1. Introduction

A six-month-old infant who has not yet acquired a single coherent sentence will, given the opportunity, begin levering herself upright against the bars of her crib. By the time she utters her first recognizable word, she will already have integrated vestibular feedback, distributed proprioception (the body's internal sense of the position and motion of its own limbs), contact mechanics on heterogeneous surfaces, online estimation of inertial parameters, and a hierarchical motor plan that anticipates center-of-mass excursions she has never before encountered. No present artificial system, regardless of its parameter count or training corpus, comes within an order of magnitude of this competence. That asymmetry is the central provocation of this review.

Contemporary large language models (LLMs) write fluent code, summarize dense literature, pass national medical board examinations, and increasingly interpolate plausible clinical reasoning [1, 2]. Yet the same systems, when coupled to the most capable bipedal hardware available in 2026, still struggle to traverse an unfamiliar curb. This is the modern face of Moravec's paradox [3]: the cognitive performances we historically considered hardest—abstract reasoning, formal logic, encyclopedic recall—have proven the most computationally tractable, while the perceptual-motor performances we share with infants, birds, and dogs remain, by a wide margin, the least.

The standard response to Moravec's paradox in the LLM era has been ambition by extrapolation: scale, more text, longer context windows, more chain-of-thought. The dissenting position—articulated most forcefully by LeCun [4] and Assran et al. [5, 6], echoed by Fei-Fei Li's call to elevate spatial intelligence to a first-class research target [7], and increasingly visible in the world-model programs of Google DeepMind [8, 9] and the generalist robot-policy work of Physical Intelligence [10]—is that the bottleneck is structural rather than quantitative. No volume of text, however large, encodes the constraints that gravity, friction, and contact dynamics impose on a body. Those constraints must be encountered, predicted, and respected; they cannot be paraphrased.

This review develops that dissenting position from the perspective of climbing and walking robotics. Three claims structure the argument. First, that the limit on present artificial general intelligence (AGI) is Newtonian in a precise sense: the failure mode is not a deficit of symbol manipulation but a deficit of grounded, predictive engagement with the physics of macroscopic bodies. Second, that the architectural responses now most credible—joint-embedding predictive architectures (JEPAs)

***Corresponding author:** Enrique Díaz Cantón, Department of Medical Oncology and Artificial Intelligence, School of Medicine, CEMIC University Institute, Argentina. Email: ediazcanton@iuc.edu.ar

and interactive world models—converge on a research agenda for which legged and climbing robots are uniquely well-suited testbeds. Third, that locomotion deserves to be treated not as one engineering problem among many but as a privileged benchmark for embodied cognition, occupying a role analogous to the one we have proposed elsewhere for elite surgery [11] and competitive Brazilian Jiu-Jitsu [12]: an upper-bound, integrative challenge whose solution would entail solving most of the harder substructure of embodied AGI. We stress at the outset that these three claims are advanced as falsifiable hypotheses rather than established results: §1.1 makes the review's methodology explicit, §3.1 weighs the principal rival positions, and §7 specifies the evidence that would confirm or refute them.

The review is organized as follows. Section 2 introduces a Pauline diagnostic—the vivimus/movemur/sumus triad—as a heuristic for locating where artificial systems currently stand and where they remain absent (Figure 1). Section 3 examines why language alone plateaus for embodied agency. Section 4 develops the Newtonian framing and the case for predictive world models, drawing on JEPA, V-JEPA 2, and the Genie line of work. Section 5 argues that locomotion (broadly construed to include climbing, dynamic walking, and stair- and rubble-traversal) is the privileged site at which the next generation of embodied AGI ought to be tested. Section 6 reviews the recent empirical landmarks of learning-based legged locomotion (2019–2026). Section 7 lays out a milestone ladder (Table 3). Section 8 enumerates technical requirements. Section 9 addresses the sim-to-real gap. Section 10 closes with the philosophical residue and the conclusion.

1.1. Scope and review methodology

This article is a narrative (perspective) review rather than a systematic review; it does not follow a PRISMA protocol and makes no claim to exhaustive coverage. Its aim is to synthesize and interpret converging developments across three literatures—embodied cognition, predictive world models, and learning-based legged and climbing locomotion—and to articulate, from that synthesis, a falsifiable thesis about the embodied bottleneck of AGI.

Sources were identified by searching arXiv, IEEE Xplore, Science Robotics, and Google Scholar for the machine-learning and robotics literature, and PubMed and PhilPapers for the cognitive-science and philosophy-of-mind literature, using combinations of the terms "legged locomotion learning," "quadruped/humanoid reinforcement learning," "world models," "joint-embedding predictive architecture," "sim-to-real transfer," "embodied cognition," and "symbol grounding." Empirical landmarks (Section 6) were prioritized for the period 2019–2026, with foundational earlier works retained on grounds of conceptual relevance. Inclusion of an empirical result required peer-reviewed publication or a high-visibility preprint, a demonstration on physical hardware or at established simulation benchmarks, and salience within the legged-locomotion community as judged by citation and replication.

Two limitations follow from this design and are stated explicitly. First, the selection of landmarks reflects the authors' interpretive judgment and is offered as a reasoned synthesis rather than a complete census of a fast-moving field. Second, because the review advances a thesis, care has been taken throughout to distinguish empirically established findings from the interpretive claims built upon them; the milestone ladder of Section 7 is the mechanism by which those interpretive claims are rendered falsifiable.

2. A Diagnostic Triad: Vivimus, Movemur, Sumus

We have argued elsewhere [11] that the Pauline formula in the Acts of the Apostles—In ipso vivimus et movemur et sumus, "in him we live, and move, and have our being" (Acts 17:28)—provides a heuristic, almost diagnostic, mapping of the present trajectory of artificial intelligence (AI). The triad partitions the territory cleanly (Figure 1; Table 1). Vivimus, read

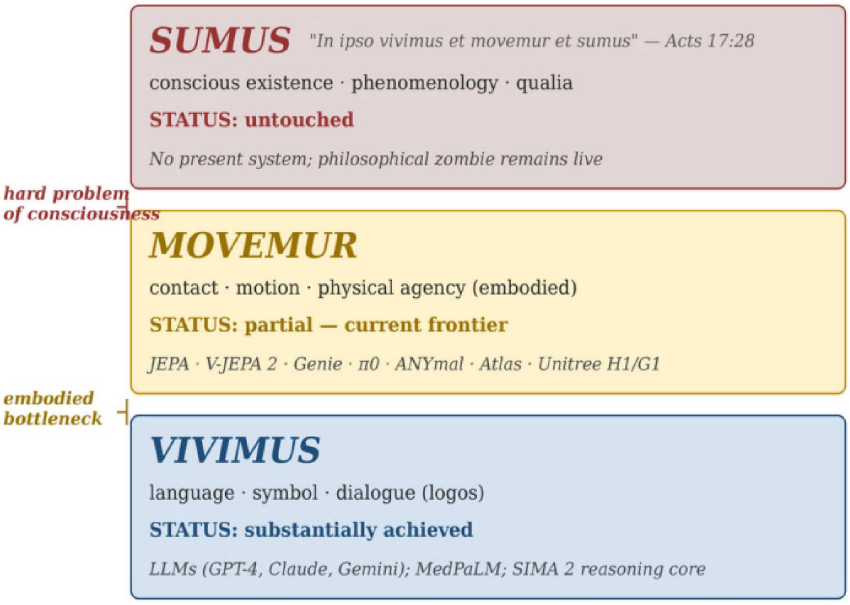

Figure 1
The vivimus/movemur/sumus triad as a diagnostic for artificial intelligence. Linguistic competence (vivimus) is substantially achieved; embodied competence (movemur) is the current frontier; conscious existence (sumus) remains untouched. The embodied bottleneck lies between the first two layers, the hard problem of consciousness between the last two

narrowly as the domain of logos (the Greek term for word, reason, and discourse)—language, dialog, symbol—has been substantially conquered. LLMs do not merely simulate fluent linguistic exchange; they sustain it across registers, languages, and domains in ways that, a decade ago, were considered indefinitely distant. By any reasonable behavioral test of linguistic competence, the vivimus threshold has been crossed [1]. In plain terms, vivimus asks whether a system can handle language; movemur, whether it can act competently in the physical world; and sumus, whether it has conscious experience.

The movemur—the domain of motion, of bodily engagement with the physical world—is a different country. Here, the most advanced systems available at the time of writing produce results that are, frankly, partial. Quadruped platforms now traverse rough terrain robustly [13, 14]; humanoid platforms execute scripted manipulation tasks [15]; vision-language-action models such as $\pi 0$ [10] generalize across robot embodiments to a degree unimaginable five years ago. But the tasks at which infants effortlessly outperform machines remain numerous: stair-climbing on novel staircases, recovery from an unexpected slip on ice, balancing on a moving boat, carrying a bowl of soup across uneven ground without spilling. The distance between vivimus and movemur is not an incremental gap; it is a structural one and the principal subject of this review.

Table 1
Diagnostic mapping of artificial intelligence onto the Pauline triad vivimus/movemur/sumus

Triad term	Cognitive domain	Current AI status	Representative systems/ references
Vivimus	Language, dialog, symbolic reasoning (logos)	Substantially achieved	LLMs (GPT-4, Claude, Gemini); MedPaLM [1]; SIMA 2 reasoning core [9]
Movemur	Contact, motion, physical agency under Newtonian constraints	Partial; current research frontier	JEPA/V-JEPA 2 [4, 6]; Genie [8]; $\pi 0$ [10]; ANYmal, Atlas, Unitree H1/G1 [15]
Sumus	Conscious existence, phenomenology, qualia	Untouched	No present system; philosophical-zombie possibility remains live [12]

The sumus—the dimension of conscious existence, of phenomenology, of qualia—is, of course, untouched, and this review does not pretend to address it. We have written on this question elsewhere [12], and the position there developed remains the operative one: even an artificial system that reliably traverses the movemur threshold may still be, in David Chalmers's vocabulary, a philosophical zombie. The cautionary inference is symmetrical: solving locomotion does not entail solving consciousness, and we should not let the engineering achievement of the former obscure the philosophical opacity of the latter.

What this review claims, then, is more modest and more specific than an AGI-completeness argument. It claims that movemur is the immediate frontier, that it is the limit on which contemporary architectures most plausibly fail, and that a research program centered on legged, climbing, and contact-rich robotic agents is therefore not a peripheral specialty within AI but a candidate site for the next decisive breakthrough.

Table 1 expands the diagnostic mapping with representative systems and status assessment, summarizing the conceptual content of Figure 1 in tabular form.

3. Why Language Alone Plateaus for Embodied Agency

Before turning to architectures, this section explains in plain terms why simply making language models larger does not, on its own, produce physical competence. The plateau argument has both an empirical and a structural form. Empirically, scaling laws for language modeling, robust though they remain on text-prediction benchmarks, have not translated into corresponding gains in physical task competence. A model trained on the entire indexed web is, on locomotion, no closer to an infant than a model trained on a small fraction of it. Structurally, the reason is straightforward: text records the propositional residue of human cognition, but very little of human cognition is propositional. Lakoff and Johnson [16], Varela et al. [17], and Brooks's earlier polemics [18, 19] converge on this: a great deal of what we call understanding is sedimented from continuous bodily engagement with the world and is never linguistically articulated, because it does not need to be.

Harnad's symbol-grounding problem [20] posed the question with surgical clarity: how do symbols acquire meaning? The behavior of contemporary LLMs is not so much a refutation of Harnad's problem as a vivid instantiation of it. The systems manipulate tokens with breathtaking facility; the tokens, however, remain ungrounded in any physical referent. This becomes plainly visible the moment one couples an LLM to actuators. Dialog confidence does not survive contact with a friction discontinuity.

There is a complementary structural point. Contact-rich tasks demand four properties simultaneously: (i) high-bandwidth, multimodal sensing—vision, proprioception, force, tactile and inertial—at sample rates that text generation never approaches; (ii) low-latency control under irreducible uncertainty, with reflex-like responses on millisecond timescales; (iii) accurate prediction of how interventions will alter the state of the physical world; and (iv) hierarchical planning that interleaves these timescales coherently. None of these properties is reducible to next-token prediction over discrete vocabularies. They are, in LeCun's phrase, capacities that models of the world must support directly [4]. Recent comprehensive reviews of embodied AI [15] underscore the same conclusion from an engineering vantage: cyberspace and the physical world are not aligned by parameter count alone.

3.1. Alternative positions and objections

Intellectual honesty requires that the thesis of this review be weighed against the strongest rival positions. We identify three, and in each case state what evidence would count against our own view.

The first is the scaling-maximalist position: that sufficiently large multimodal models—trained on enough text, image, and video—will acquire adequate internal models of the physical world without ever being embodied, so that embodiment becomes

a convenience rather than a necessity. This view draws real support from evidence that predictive sequence models can develop structured internal representations of the domains on which they are trained; the most striking demonstration is that a transformer trained only to predict legal moves in a board game nonetheless learns an internal representation of the board state it was never explicitly given [21], alongside the broader observation that qualitatively new capabilities can emerge with scale [22]. We take this objection seriously. Our reply is that representing the state of a discrete, fully observed game is not the same problem as predicting continuous contact dynamics under partial observation and irreversible consequence—the competence that the intuitive-physics and core-knowledge literatures locate at the root of human cognition [23, 24]—and that the empirical record still shows no transfer from linguistic scale to physical competence.

The second is the simulation-is-sufficient position: that learned or simulated virtual worlds (Section 4.3) can substitute entirely for physical bodies, making real robots unnecessary for the development of embodied intelligence. We regard simulation—and the learned world-model program that extends it [25, 26]—as indispensable but not sufficient, for the reason developed in Section 9: the residual sim-to-real gap is precisely where the unmodeled physics that defeats deployed systems resides.

The third is a deflationary objection to the notion of embodiment itself—that "embodied" competence is merely a heterogeneous collection of engineering problems with no privileged core, so that singling out locomotion is arbitrary. Our response, developed in Section 5, is that locomotion is distinguished not by fiat but by its integrative structure: it compresses balance, contact, prediction, and hierarchical control into a single continuous task under unforgiving physical constraint.

We therefore state the thesis in falsifiable form. If a disembodied model, trained only on text and passive video, were to control a novel legged platform across unfamiliar natural terrain at the level of Section 7's L4–L5 milestones, the central claim of this review would be substantially weakened. That this experiment has not succeeded, despite the scale now available, is the strongest evidence for the position we defend.

4. Newtonian Physics and the Case for Predictive World Models

This section states precisely what we mean by a "Newtonian" limit and why predictive world models are the natural response to it.

4.1. What "Newtonian" means here

By Newtonian physics, we do not mean the historical theory in its narrowest sense—Newton's three laws and universal gravitation as expounded in the Principia. We mean, instead, the macroscopic, classical-mechanical regime within which biological bodies operate: rigid- and soft-body dynamics, gravity, friction, contact, momentum, deformable interfaces, fluid drag, and the energetic constraints of locomotion. This is the physics that any animal, and any robot that aspires to function alongside animals, must respect on every timestep. It is also the physics that no amount of textual training data can substitute for because the relevant constraints are continuous, online, and unforgiving. A simulation that violates them does so at the cost of a fall.

The phrase "folk physics" or "intuitive physics" is sometimes used in this connection. We prefer Newtonian, with the qualifications above, because the term resists the inference that what infants do is in some sense informal or approximate. It is not. The infant's world model is informally articulated, but it is, as a predictive engine, often considerably more reliable on the relevant timescales than any explicit model the field of robotics currently deploys.

4.2. LeCun's JEPA and the move beyond pixel-perfect generation

LeCun's [4] proposal for autonomous machine intelligence is, in essence, an architectural argument that predictive world models built on joint-embedding objectives will outperform both supervised and generative-pixel approaches for embodied tasks. The relevant insight, formalized in I-JEPA [5] and its successors, is that an agent need not predict the future at the level of pixels. It needs, rather, to predict the future in a representation space that captures task-relevant latent structure. Pixel prediction is wasteful because most of the bits in an image are irrelevant to action; latent prediction is parsimonious because it allocates representational capacity to what matters.

For a climbing or walking robot, the JEPA family of architectures offers something specific: a path to agents that can mentally rehearse the consequences of a foothold, a handgrip, or a balance recovery before committing to it irreversibly and that can do so without the brittleness associated with photorealistic generative simulation. The relevant predictions concern center-of-mass trajectories, contact-force distributions, slip likelihoods, and stability margins—all of which live naturally in latent representation, not in pixel space.

4.3. DeepMind's genie and interactive learned worlds

If JEPA addresses the question of how a world model should be structured, the Genie line of work [8] addresses the orthogonal question of where the data comes from. Genie demonstrates that interactive, action-controllable virtual environments can be learned from large corpora of unlabeled video. The implications for embodied AI are direct: the cost of generating training environments for legged or climbing agents—historically the bottleneck of sim-to-real—can in principle be amortized across the immense quantity of human-generated visual data already available.

The SIMA program and its SIMA 2 extension [9] add a complementary thread: scaling instructable generalist agents across many simulated worlds, with a generic humanlike interface, and increasingly ambitious multistep goal completion. The Genie–SIMA pairing—learned worlds plus generalist agents—sketches a credible path to training embodied policies at scale on diversity unattainable from physical robot data alone.

4.4. V-JEPA 2 and the consolidation of the world-model program

Two empirical developments published since the original conception of this argument deserve explicit mention. First, V-JEPA 2 [6] extends the joint-embedding predictive principle to video, pre-trained on more than one million hours of internet footage and demonstrating state-of-the-art performance on motion understanding and human-action anticipation, while supporting zero-shot model-based planning on real robots with only modest interaction data. This is the strongest existence proof to date that LeCun's architectural prescription is not merely programmatic but empirically actionable. Second, the

cross-embodiment policy work of Physical Intelligence [10] has shown that a single vision-language-action policy can transfer across heterogeneous robot bodies, mitigating the long-standing one-body, one-policy bottleneck that has slowed the deployment of learned controllers to legged platforms.

The convergence of these lines of work—JEPA's representational discipline, V-JEPA 2's video-scale realization [6], Genie's data leverage, SIMA's behavioral generality, $\pi 0$'s cross-embodiment policy training, and even the existence proof from large-scale physical prediction in adjacent domains such as weather forecasting [27]—suggests that the architectural pieces are becoming available. What remains is the integration, and the integration must be tested on bodies. Bodies do not abstract. Table 2 summarizes the convergent architectures and their respective roles in this research program.

5. Walking and Climbing as Privileged Testbeds for Embodied AGI

Within the broader landscape of embodied benchmarks, locomotion in general—and climbing in particular—has properties that single it out as a privileged testbed.

First, locomotion is integrative. It compresses into a single, continuous task most of the requirements for embodied intelligence: dynamic balance, contact mechanics on heterogeneous surfaces, online inertial estimation, hierarchical motor planning across millisecond and second timescales, anticipation of self-induced perturbations, and recovery from external ones. Solving locomotion well does not solve embodied intelligence, but it cannot be solved in isolation—its substructure is the substructure of embodiment itself [28].

Second, locomotion is unforgiving. Failure modes are immediate, often irreversible, and physically legible. A mis-stepping LLM fabricates a plausible footnote; a mis-stepping biped falls. The cost gradient sharpens evaluation: there is no equivalent of confabulation in the locomotion regime. Either the agent stays upright, or it does not. This property, far from being a disadvantage for AGI evaluation, is one of locomotion's distinctive strengths. It exposes failure modes that text benchmarks systematically obscure.

Third, locomotion is ecologically grounded. Animal locomotion has been refined under selective pressure for hundreds of millions of years, and the resulting design constraints are encoded in physiology with an economy that bears study. The energetic cost of transport in efficient quadrupeds and bipeds approaches the thermodynamic minimum for terrestrial movement; balance recovery in elite human movers exhibits anticipatory postural adjustments on the order of 100 ms before a perturbation arrives. Any artificial system that aspires to general embodied competence in unconstrained environments must approach these baselines, and this places a non-arbitrary, biologically calibrated bar on the field.

Fourth, climbing—distinctively the focus of this journal and its associated community—adds requirements that flat-ground walking does not impose: three-dimensional planning, reach-and-grasp coordination, force distribution across noncoplanar contacts, planning under variable hold quality, and risk management under non-Markovian failure topologies. Recent work on intelligent rock-climbing robots with hybrid bioinspired attachment mechanisms [29] illustrates the level of integration required: a single substrate combining gripper morphology, contact-aware control, and multimodal locomotion. A robot that can climb a nontrivial natural feature (an irregular boulder, a tree, a partially collapsed structure) must integrate, on a single substrate, contact-rich manipulation and contact-rich locomotion. There are very few tasks in robotics that demand this integration as forcefully. Climbing, in this sense, is to embodied AI what elite surgery is to medical AI [11] and what high-level grappling is to humanoid combat AI [12]: an upper-bound, integrative benchmark that exposes, simultaneously, almost every relevant deficit.

Table 2
Convergent world-model and embodied-agent architectures relevant to legged and climbing robotics

System/family	Originator (year)	Architectural contribution	Relevance to climbing/walking robotics
JEPA/I-JEPA	Meta FAIR [4, 5]	Joint-embedding predictive learning in latent space; non-generative	Latent rehearsal of footholds, balance recoveries, contact transitions
V-JEPA 2	Meta FAIR [6]	Video JEPA pre-trained on >1 M h of internet video; supports zero-shot model-based planning on real robots	Strongest existence proof that JEPA scales to dynamic, contact-relevant prediction
Genie	Google DeepMind [8]	Generative interactive environments learned from unlabeled video	Cheap synthesis of diverse training environments; alleviates sim-to-real data bottleneck
SIMA/SIMA 2	Google DeepMind [9]	Generalist instructable agent across many simulated 3-D worlds; Gemini-based reasoning core	Demonstrates behavioral generality and self-improvement across embodied virtual tasks
π0	Physical Intelligence [10]	Vision-language-action flow model; cross-embodiment robot control	Path to single-policy generalist controllers usable across legged and manipulating platforms
GraphCast	Google DeepMind [27]	Large learned predictive model of atmospheric dynamics	Existence proof that high-dimensional physics is learnable end to end; suggestive analogy for whole-body dynamics

6. Recent Empirical Landmarks in Legged Locomotion (2019–2026)

Any review aimed at the climbing-and-walking-robots community would be incomplete without an explicit synthesis of the empirical landmarks that have reshaped legged-robot learning over the last seven years. We highlight five that, in our reading, are constitutive of the present state of the art.

Hwangbo et al. [30], publishing in Science Robotics, demonstrated that deep-reinforcement-learning policies trained entirely in simulation could be transferred zero-shot to the ANYmal quadruped, achieving agile motor skills previously unattainable with classical state-machine controllers. Lee et al. [13], in the same journal, then showed that those same learned policies retained their robustness in genuinely challenging natural terrain—mud, snow, vegetation, gushing water—relying only on proprioceptive feedback. Miki et al. [14] closed the perception–locomotion loop, integrating exteroceptive depth perception with proprioceptive control to produce a system capable of operating in the wild over long durations. These three papers, taken together, constitute the empirical demonstration that the learning-based program for quadruped locomotion is not aspirational but executable.

Cheng et al. [31] extended this trajectory to "extreme parkour," training a single-legged robot policy to leap, climb, and traverse obstacles that exceed the platform's nominal dimensions, using egocentric vision. The implication is that the integrative capacities long demanded by climbing-robot researchers—multistep planning, contact-aware control, anticipation across timescales—are emerging within the same architectural paradigm as routine quadruped walking. The boundary between "walking robot" and "climbing robot" is, on a learning-architecture analysis, less firm than it once appeared.

On the humanoid side, the period 2023–2025 saw the public demonstration of bipedal platforms—Boston Dynamics' Atlas, Tesla Optimus, Unitree H1 and G1, Figure 02, Agility Digit—capable of stable locomotion, stair-climbing, and basic loco-manipulation, with control policies increasingly trained end to end rather than scripted [15, 32]. The comprehensive review of bipedal walking robots by Mikolajczyk et al. [28] provides the longer historical baseline against which this acceleration is most legible. Across all of these systems, a recurring lesson emerges: reliable performance in the world depends less on any one breakthrough than on the joint maturation of simulators, learning algorithms, hardware compliance, and tactile sensing.

Distributed tactile sensing for legged platforms is itself an active frontier. Recent work on soft-surfaced, vision-based tactile sensing for biped feet [33] demonstrates that contact-state estimation—position, orientation, shear, center of pressure (the point on the foot where the ground's reaction force effectively acts), terrain classification—can be obtained directly from foot-borne haptic signals, with measurable improvements in balance control. This is the engineering analog of the proprioceptive substrate that animals rely on, and its maturation is a precondition for the higher reaches of the milestone ladder presented below.

Beyond quadrupedal walking, the same learning-based paradigm has now produced peer-reviewed hardware results across a widening range of regimes: agile obstacle "parkour" navigation on the ANYmal quadruped [34], learned locomotion on genuinely deformable substrates such as sand [35], and fully learning-based real-world humanoid walking driven by a causal-transformer policy [36]. At the edge of physical feasibility, champion-level autonomous drone racing [37] offers a further existence proof that end-to-end learned control can match expert human performance in a fast, unforgiving physical task. These results collectively strengthen—without yet completing—the empirical case that the learning-based program scales from flat-ground walking toward the integrative demands of climbing.

7. A Milestone Ladder for Locomotion-Grounded Embodied AGI

A benchmark cannot be useful if it is binary. Following the milestone-ladder format, we have proposed for related embodied-AGI testbeds [11, 12], we outline below a graded sequence of capability levels for locomotion-grounded embodied agents (Table 3). The ladder is intended to be evaluable, falsifiable, and progressively integrative; it does not assume that levels are crossed in strict order, but it does assume that no level above L4 is credibly approached without robust performance at the levels below it. By the criterion implicit in this ladder, no contemporary system is above Level 3, and most are at Levels 1–2 with significant caveats.

Level 6, as described, is—by any honest assessment—beyond the present state of the art. It is offered not as a near-term roadmap but as the upper bound that the field's progress should be measured against.

8. Technical Requirements for Contact-Rich Embodied Intelligence

Achieving even intermediate milestones on the above ladder demands integration of perception, control, and learning under irreducible uncertainty. The following five requirements are, we argue, jointly necessary.

8.1. Distributed sensing and embodiment

Robust embodied agents require distributed tactile sensing across feet, hands, and contact-prone surfaces; force–torque transducers at major joints; high-fidelity proprioception; and calibrated multimodal vision (stereoscopic, depth, and event-based where feasible). A walking or climbing agent that lacks force feedback at its contacts is, in a precise sense, deaf to the most informative signals about its own state. The recent demonstration of soft-surfaced vision-based tactile sensing for biped feet [33] confirms that the relevant signal is now technically retrievable; the engineering remit is to make it standard [38]. Classical observation—that the structure of the environment imposes constraints on what representations a perceiving system must compute—applies directly: the relevant information for locomotion lives largely in contact, not in pixels.

8.2. Compliant control and reflex-like behaviors

Low-latency impedance control (controlling how a limb yields to force, rather than commanding its position alone), mechanically compliant actuation (whether through series-elastic actuators, variable-stiffness actuators, or soft-robotic substrates), and reflex-like protective behaviors operating on millisecond timescales are not optional features for safe contact-rich operation. They are the engineering analog of the spinal reflex arcs that allow animals to recover from perturbations on timescales faster than supraspinal control can manage. An agent that cannot reflexively absorb a missed step or a slipping handhold cannot, in any robust sense, be said to walk or climb.

Table 3
Milestone ladder for locomotion-grounded embodied artificial general intelligence

Level	Capability	Operational definition	State of the art (2026)
L0	Static stability and posture control	Maintenance of upright posture under quasi-static conditions on flat, rigid ground; recovery from small perturbations within support polygon	Solved
L1	Periodic locomotion on engineered terrain	Steady walking, trotting, or climbing on flat, regular, engineered surfaces with known friction; energetic cost benchmarked against animal baselines	Solved [30]
L2	Reactive locomotion under bounded disturbances	Balance recovery from external pushes within prescribed force envelopes; navigation of small scripted obstacles; transitions between gaits	Substantially solved
L3	Adaptive locomotion on heterogeneous natural terrain	Robust traversal of unfamiliar but accessible natural surfaces: grass, gravel, light rubble, mild slopes, low stairs. Online estimation of friction and compliance	Achieved [13, 14]
L4	Climbing on structured natural and built features	Reliable ascent and descent of staircases, ladders, and engineered climbing structures; basic hand-and-foothold recognition; dynamic recovery from a missed grip	Emerging [29, 31]
L5	Climbing under uncertainty, with task integration	Ascent of irregular natural features under uncertain hold quality; integration with manipulation goals; recovery from low-probability failures without external assistance	Open
L6	Generalist embodied competence across loco-manipulation	Full L0–L5 spectrum on a single substrate, with online transfer between locomotion regimes, graceful degradation under faults, and reliable cooperation with humans in shared workspaces	Open; beyond present state of the art

8.3. Hierarchical planning across timescales

Expert human movers operate simultaneously at three temporal scales: millisecond reflexes, second-level motor primitives, and minute-to-hour strategic planning. Autonomous embodied agents will require analogous layered architectures—reactive stabilization loops, mid-level skill modules (gait initiation, foot placement selection, hold acquisition), and high-level planners capable of revising goals as terrain unfolds. The architectural challenge is not merely to instantiate each layer but to ensure coherent communication and override authority between them.

8.4. Predictive world models

This is the architectural locus of the JEPA argument [4–6]. An embodied agent that can internally simulate the consequences of a contemplated action—"if I place my left hand here, will the center of mass remain within the new support polygon (the base of support outlined by the feet in contact with the ground)?"—possesses a faculty that pure reactive policies and pure model-free reinforcement learning cannot easily acquire. Predictive world models are also the natural substrate for transfer between simulated and physical regimes because the latent representations they learn are, by construction, less brittle than pixel-perfect generative ones. V-JEPA 2 [6] provides the first compelling demonstration that this faculty is achievable from internet-scale video, and the Genie line of work [8] complements it by providing the data infrastructure for training such models at scale.

8.5. Hardware-aware safe learning

Embodied agents differ from disembodied ones in that the cost of an erroneous action is paid in physical hardware, possibly in injury to humans, and always in time. Safe learning under irreversibility is therefore not a deployment concern bolted onto a research program; it is a constitutive feature of the program. The relevant techniques—constrained reinforcement learning, shielding, formal-method-backed runtime monitors, and rigorous human-in-the-loop oversight during the upper levels of the milestone ladder—must mature in parallel with the architectural and perceptual advances above.

9. Sim-to-Real and the Asymmetric Cost of Being Wrong

In plain terms: a mistaken word can simply be edited, but a mistaken movement can break the robot or injure a bystander, and that asymmetry shapes the discussion below. High-fidelity

physics engines [39] and large-scale parallel simulation accelerate algorithmic iteration on legged and climbing agents in ways that physical experimentation alone cannot match. They are, however, no substitute for the reality they model. The classical sim-to-real problems—discrepancy in friction modeling, soft-tissue and soft-surface compliance, sensor noise spectra, latency distributions, and the long tail of unmodeled contact regimes—are, in practice, the dominant source of failure in deployed systems [40].

Three responses are, individually, incomplete and, in combination, plausibly sufficient. Aggressive domain randomization (deliberately varying the simulator's physics and appearance so the trained policy transfers to the real world) distributes the burden of robustness across simulator parameters that would otherwise be tuned to a single regime. Careful system identification narrows the simulator-to-physical gap on parameters that are unfavorable to randomize. Structured real-world data collection—increasingly tractable thanks to cross-embodiment policy learning [10] and to learned world models that can ingest unlabeled video [6, 8]—closes the residual gap by exposing policies to the full distribution of physical surprises. The recent survey of sim-to-real transfer methods in deep reinforcement learning [40] provides a useful taxonomy of these strategies and their respective trade-offs.

An asymmetry deserves explicit naming. In the linguistic regime, the cost of being wrong is reputational and recoverable. In the embodied regime, it is structural and frequently irrecoverable. A robot that falls from a structural failure in its world model damages itself, possibly damages its surroundings, and possibly damages a human bystander. This asymmetry argues for a research culture in which conservative deployment, rigorous staged evaluation along the milestone ladder of Section 7, and a strong epistemic preference for falsifiable benchmarks are treated as constitutive professional norms—analogous to those long established in surgical and clinical research [11]. The robotics community has, on this front, much to learn from, and to contribute to, medicine. The foundational technique of domain randomization [41] remains central to these strategies.

10. Discussion and Conclusion: Movemur, Sumus, and the Limits of the Argument

Three caveats are worth stating plainly. First, this review does not argue that solving locomotion solves AGI. It argues that not solving locomotion makes claims of AGI rhetorical. The conditions are necessary but not sufficient. The embodied bottleneck is real; clearing it does not make the linguistic, social, and phenomenological bottlenecks disappear.

Second, the architectural prescriptions outlined here—JEPA-style world models, V-JEPA 2 video-scale realizations, Genie-style learned environments, generalist vision-language-action policies, hierarchical compliant control—are credible candidates, not validated solutions. The history of AI is unkind to architectural overconfidence, and the present review aspires to identify the right class of approach rather than to anoint any specific instance within it. The empirical work of the next decade will adjudicate.

Third, even when movemur is solved—when an autonomous system reliably climbs an unfamiliar boulder, recovers from an unanticipated slip, and traverses rubble with a reasonable energy budget—the resulting agent will not, on that ground alone, possess sumus. It may be a philosophical zombie of the most accomplished sort: behaviorally indistinguishable from an embodied conscious agent in the physical regimes for which it was trained, and yet, ex hypothesi, lacking phenomenological interiority [12]. The right inference from this is not despair about the AGI project but disciplined humility about the difference between behavioral competence and conscious existence. Walking is not having; movemur is not sumus.

What this review does claim, with some confidence, is that the climbing and walking robotics community is positioned to contribute to the next decisive advance in AGI not as a downstream consumer of architectures developed elsewhere but as a primary site at which those architectures are tested, falsified, and refined. The interface between learned world models and physical bodies is not somebody else's problem. It is the problem.

The thesis of this review has been straightforward and, we hope, directly actionable for the JCWR community. Intelligence that operates in the physical world is constrained by physics, contact, and consequence. Linguistic competence—vivimus—has been substantially achieved, but it has not generalized, and there is principled reason to expect that it will not generalize, to the embodied competence—movemur—that biological agents acquire as a matter of course. The architectural responses now most credible converge on predictive world models, joint-embedding representations, and learned interactive environments; their natural testbed is the contact-rich, gravity-bound, irreversibly consequential regime in which legged and climbing robots operate.

The day an autonomous agent reliably ascends an unfamiliar staircase, recovers from a slip on ice, and carries a hot beverage across uneven ground without spilling will mark not merely a robotics achievement but a substantive expansion of what the term AI has earned. Until that day, walking and climbing robots are not adjacent to the AGI question. They are where it is decided. We offer this closing formulation as a deliberately strong statement of the review's thesis, not as an empirical finding: as emphasized throughout, solving locomotion is argued to be necessary but not sufficient for embodied general intelligence.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Enrique Díaz Cantón: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., . . . , & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2

[2] Truhn, D., Reis-Filho, J. S., & Kather, J. N. (2023). Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine*, *29*(12), 2983–2984. https://doi.org/10.1038/s41591-023-02594-z

[3] Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. USA: Harvard University Press.

[4] LeCun, Y. (2022). *A path towards autonomous machine intelligence*. *OpenReview*.

[5] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., . . . , & Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629. https://doi.org/10.1109/CVPR52729.2023.01499

[6] Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., ..., & Ballas, N. (2025). V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv Preprint: 2506.09985.*

[7] Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L., & Xie, S. (2025). Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10632–10643. https://doi.org/10.1109/CVPR52734.2025.00994

[8] Bruce, J., Dennis, M., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., ..., & Rocktäschel, T. (2024). Genie: Generative interactive environments. *Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235*, 4603–4623.

[9] SIMA Team. (2025). SIMA 2: A generalist embodied agent for virtual worlds. *arXiv Preprint:2512.04797*.

[10] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., et al. (2025). π0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems (RSS)*.

[11] Díaz Cantón, E. (2026). 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 an upper-limit benchmark for embodied AGI (Artificial General Intelligence)]. *Revista Argentina de Cirugía*, *118*(1), 1.

[12] Díaz Cantón, E., Casarini, I., & Díaz Cantón, M. (2026). From Res Cogitans to Digital Twins: Large Language Models, Large World Models, and the illusion of consciousness in the Brazilian Jiu-Jitsu Singularity. *Iberoamerican Journal of Medicine*, *8*(2), 32–33. https://doi.org/10.53986/ibjm.2026.0006

[13] Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., & Hutter, M. (2020). Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, *5*(47), eabc5986. https://doi.org/10.1126/scirobotics.abc5986

[14] Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., & Hutter, M. (2022). Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, *7*(62), eabk2822. https://doi.org/10.1126/scirobotics.abk2822

[15] Sheng, Q., Zhou, Z., Li, J., Mi, X., Xiang, P., Chen, Z., ..., & Yang, H. (2025). A comprehensive review of humanoid robots. *SmartBot*, *1*(1), e12008. https://doi.org/10.1002/smb2.12008

[16] Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. USA: Basic Books.

[17] Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. USA: MIT Press.

[18] Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, *6*, 3–15. https://doi.org/10.1016/S0921-8890(05)80025-9

[19] Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

[20] Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

[21] Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. In *Proceedings of The Eleventh International Conference on Learning Representations*.

[22] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ..., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

[23] Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

[24] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

[25] Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), In *Advances in Neural Information Processing Systems, 31*.

[26] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. *Nature*, *640*(8059), 647–653. https://doi.org/10.1038/s41586-025-08744-2

[27] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ..., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

[28] Mikolajczyk, T., Mikołajewska, E., Al-Shuka, H. F. N., Malinowski, T., Kłodowski, A., Pimenov, D. Y., . . . , & Macko, M. (2022). Recent advances in bipedal walking robots: Review of gait, drive, sensors and control systems. *Sensors*, *22*(12), 4440. https://doi.org/10.3390/s22124440

[29] Zi, P., Xu, K., Chen, J., Wang, C., Zhang, T., Luo, Y., . . . , & Ding, X. (2024). Intelligent rock-climbing robot capable of multimodal locomotion and hybrid bioinspired attachment. *Advanced Science*, *11*(39), 2309058. https://doi.org/10.1002/advs.202309058

[30] Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., & Hutter, M. (2019). Learning agile and dynamic motor skills for legged robots. *Science Robotics*, *4*(26), eaau5872. https://doi.org/10.1126/scirobotics.aau5872

[31] Cheng, X., Shi, K., Agarwal, A., & Pathak, D. (2024). Extreme parkour with legged robots. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 11443–11450. https://doi.org/10.1109/ICRA57147.2024.10610200

[32] Wang, J., Wang, C., Chen, W., Dou, Q., & Chi, W. (2024). Embracing the future: The rise of humanoid robots and embodied AI. *Intelligent Robotics*, *4*, 196–199. https://doi.org/10.20517/ir.2024.12

[33] Li, H., Nam, S., Lu, Z., Yang, C., Psomopoulou, E., & Lepora, N. F. (2024). BioTacTip: A soft biomimetic optical tactile sensor for efficient 3D contact localization and 3D force estimation. *IEEE Robotics and Automation Letters*, *9*(6), 5314–5321. https://doi.org/10.1109/LRA.2024.3387111

[34] Hoeller, D., Rudin, N., Sako, D., & Hutter, M. (2024). ANYmal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics*, *9*(88), eadi7566. https://doi.org/10.1126/scirobotics.adi7566

[35] Choi, S., Ji, G., Park, J., Kim, H., Mun, J., Lee, J. H., & Hwangbo, J. (2023). Learning quadrupedal locomotion on deformable terrain. *Science Robotics*, *8*(74), eade2256. https://doi.org/10.1126/scirobotics.ade2256

[36] Radosavovic, I., Xiao, T., Zhang, B., Darrell, T., Malik, J., & Sreenath, K. (2024). Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, *9*(89), eadi9579. https://doi.org/10.1126/scirobotics.adi9579

[37] Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023). Champion-level drone racing using deep reinforcement learning. *Nature*, *620*(7976), 982–987. https://doi.org/10.1038/s41586-023-06419-4

[38] Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. https://doi.org/10.7551/mitpress/9780262514620.001.0001

[39] Todorov, E., Erez, T., & Tassa, Y. (2012). MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. https://doi.org/10.1109/IROS.2012.6386109

[40] Zhao, W., Queralta, J. P., & Westerlund, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 737–744). https://doi.org/10.1109/SSCI47803.2020.9308468

[41] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 23–30. https://doi.org/10.1109/IROS.2017.8202133

How to Cite: Díaz Cantón, E. (2026). Movemur: Large World Models, Newtonian Physics, and the Embodied Bottleneck of Artificial General Intelligence. *Journal of Climbing and Walking Robots*. https://doi.org/10.47852/bonviewJCWR620210350