COGNITIVE MODELS AS EVALUATION PRIMITIVES IN HUMANOID FACTORS

Xinyuan Liu
School of Computing and Augmented Intelligence
Arizona State University, AZ, USA
xinyua11@asu.edu

Eren Sadikoglu
School of Manufacturing Systems and Networks
Arizona State University, AZ, USA
esadikog@asu.edu

Ransalu Senanayake
School of Computing and Augmented Intelligence
Arizona State University, AZ, USA
ransalu@asu.edu

Lixiao Huang
Human Systems Engineering
Arizona State University, AZ, USA
Lixiao.Huang@asu.edu

ABSTRACT

Foundation models are expanding what humanoid robots can do, but they also widen the gap between *task completion* and *human-compatible behavior*. In human environments, people coordinate using cognitive regularities—for instance, predictable timing and legible motion—that they can anticipate. Advocating for **metrics beyond task completion**, in this paper, we introduce several aspects of the Cognitive Pillar of our broader *Humanoid Factors* (HoF) framework, a cognitively grounded perspective in which **explicit cognitive models serve as evaluation primitives** for embodied reasoning and control.

As a case study, we evaluate if a reaching task trained with a *behavior cloning* neural network on a Unitree G1 humanoid follows *Fitts' Law*, a classic human psychophysics model for motor control, relating movement time to task difficulty. In simulation, rollouts can appear acceptable under completion-centric metrics; in real-robot deployment, the expected Fitts relationship breaks, revealing hesitation and timing irregularity that standard robotics metrics often miss. Overall, these preliminary results motivate our broader goal of incorporating compact cognitive models as interpretable evaluation primitives for humanoids, with the longer-term aim of applying these principles to foundation model–driven control policies, where human-compatibility failures may be even harder to detect.
Full paper draft: https://arxiv.org/abs/2602.10069

1 INTRODUCTION

For most of human history, the design of environments and tools has been centered on a single intelligent agent—the human. Human Factors (HF) emerged to optimize safety, efficiency, and usability within this human-only design paradigm. When robots were introduced, they were typically built as specialized instruments: function dictated form, and the robot's role was tightly bound to a narrow operational context.

That assumption is now being overturned. Robots are no longer designed only as fixed, single-purpose tools; instead, we are beginning to see *generalist* humanoid robots that can flexibly perform many tasks in everyday environments. This transformation is driven by large-scale *AI foundation models*: machine learning models trained on vast and diverse datasets and adaptable to a wide range of downstream tasks and contexts (Bommasani et al., 2021). When such models are embedded in humanoid robots, they provide a common substrate for perception, reasoning, and control, enabling robots to generalize across tasks and domains and to improve through continuous robot learning at scale (Gemini Robotics Team et al., 2025). Humanoids powered by these models are no longer merely preprogrammed machines; they are becoming systems that can interpret goals, learn from experience, and adapt their behavior to different situations and user preferences.

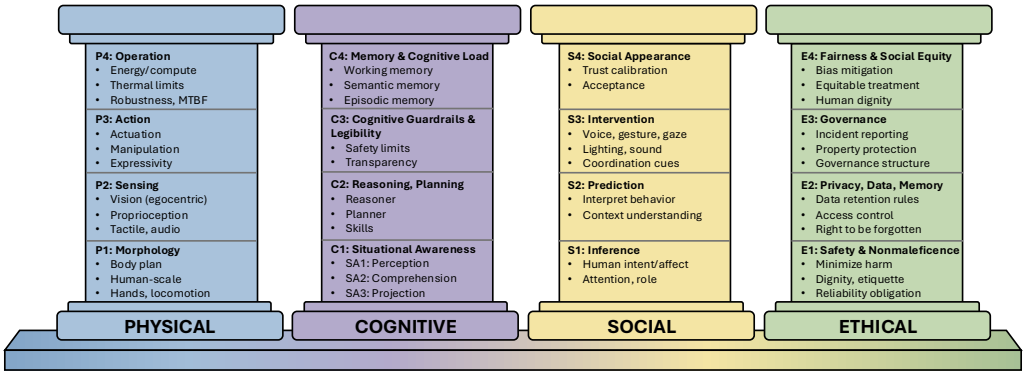


Figure 1: The Four Pillars of Humanoid Factors. This paper focuses on Layer C2.

As a result, the design challenge is no longer one-sided optimization for humans, but mutual adaptation of humans and humanoids in shared, learning-rich environments. This shift raises a fundamental question: *How should we design environments, tasks, and interaction protocols for ecosystems in which humans and humanoids coexist?*

For understanding human performance in designed systems, the field of Human Factors provides a foundation framework for analyzing how people perceive, act, and make decisions. A central principle within this field is Human-Centered Design (Xu & Gao, 2023; Shneiderman, 2020), a design philosophy and process that ensures products, systems, and services are developed with a deep understanding of users' needs, contexts, and experiences. Implicit in these approaches is the assumption that humans are the primary adaptive agents within the system.

As humanoid robots become co-actors in these same environments, the foundational goals of Human Factors remain, yet their application must broaden to encompass how both humans and humanoids jointly perceive, act, and adapt. For humanoid robots to acquire human-like sensorimotor and cognitive skills, they must continually fine-tune their underlying AI foundation models. As a result, humanoids inherit not only our physical affordances (Kemp et al., 2007)—doorways, stairs, tools—but also our social affordances (Kaufmann & Clément, 2007), such as expectations of eye contact, emotional tone, and moral reasoning. These overlapping capabilities and expectations create both opportunities for seamless collaboration and risks of misalignment.

To address this gap, we introduce Humanoid Factors, a framework that extends human-factors thinking to the design and evaluation of systems in which humanoid robots and humans interact as joint participants.

Definition. Humanoid Factors (HoF) is the design of humanoids and their surrounding environments to enable humanoids to safely, efficiently, and intelligibly adapt to and sustain operation in human-centric settings.

We therefore introduce Humanoid Factors (HoF) as the complementary counterpart to Human Factors. Whereas HF optimizes environments, tools, and workflows for human performance, HoF specifies the design, evaluation, and adaptation requirements for humanoid performance within human-built spaces. Together, HF and HoF enable design for coexistence: a mixed-agent ergonomics in which human and humanoid capabilities are specified separately (HF for humans, HoF for humanoids) and then integrated through interaction protocols, safety envelopes, and task-allocation rules. Practically, this means starting from generalist humanoids and customizing them for applications such as manufacturing, caregiving, education, and beyond.

In this view, design becomes ecological: the separation keeps accountability clear yet invites integration through a broader philosophy of design for coexistence, in which humans and humanoids operate as coordinated agents within shared environments. Generalist humanoids built on AI foundation models can serve as base platforms, customized through HoF-guided adapters, constraints, and peripherals to meet domain needs. When integrated with HF considerations, these co-adaptive

systems establish shared ergonomics (physical compatibility), shared cognition (mutual legibility of intent), shared sociality (trust and communication norms), and shared ethics (accountability and dignity in collaboration).

A growing need therefore exists to articulate a framework of Humanoid Factors that accounts for the physical, cognitive, social, and ethical dimensions shaping the design and operation of humanoid robots in human environments. Rather than positioning this as a simple extension of Human Factors, we view it as a complementary perspective that focuses on how humanoid systems—endowed with human-like embodiment and adaptive intelligence powered by foundation models—can be systematically designed, evaluated, and refined to function safely and intelligibly alongside people.

In contrast to the interaction-centric, moment-level, task-specific focus of robotics and human–robot interaction (HRI) research (Rodriguez-Guerra et al., 2021), HoF adopts a system-level, lifecycle-level, and environment-centric perspective to ask: How should a humanoid be designed so that it can safely, predictably, and acceptably live and operate in human environments? Although research on humanoid robots has advanced rapidly in the past three years (Tong et al., 2024; Cao, 2025), current literature remains fragmented across robotics (Cao, 2025), machine learning (Xiao et al., 2025), and ergonomics (Kóczi & Sárosi, 2025), with limited guidance on consistent principles for humanoid form, behavior, and integration in shared environments. We organize HoF into four separate pillars because the main challenges of humanoid coexistence arise from four distinct but interacting sources: embodiment in human-built spaces (Physical), internal reasoning and decision processes (Cognitive), coordination with human expectations and communication norms (Social), and questions of safety, responsibility, and acceptable conduct (Ethical). This paper therefore develops the concept of Humanoid Factors as a dedicated framework to consolidate these perspectives and to provide a foundation for systematic study, evaluation, and co-design of humanoids and the environments they inhabit. To make this goal actionable, we next define *Humanoid Factors* and organize it into a small set of core dimensions that can guide systematic evaluation and co-design.

2 HUMANOID FACTORS

As humanoid robots enter human spaces, design must shift to a dual-agent, co-adaptive view in which humans and humanoids learn each other's capabilities, limitations, and behavioral regularities over time (Nikolaidis & Shah, 2013; Ajoudani et al., 2018). This motivates a complementary lens: *Humanoid Factors*, the systematic study of how humanoid embodiment, intelligence, and social presence shape performance in human-designed environments, and how these factors must be understood to enable effective human–humanoid coexistence (Dautenhahn, 2007; Fong et al., 2003).

This need arises because humanoids differ fundamentally from conventional automation. Unlike purpose-built machines designed for specific tasks, humanoids combine human-like morphology with increasingly general-purpose intelligence, creating three fundamental shifts:

1. **From automation to embodiment.** Unlike conventional automation, humanoids inherit human environmental affordances by aligning with human scale, reach, and signaling. This alignment allows them to leverage existing tools and workflows, assume dull, dirty, and dangerous tasks, and collaborate more naturally and safely with people (Takayama et al., 2008). From an AI perspective, embodiment in human environments also mitigates data scarcity in training AI foundation models: humanoids can leverage large volumes of human activity recordings and demonstrations, which are largely unavailable to most other robotic platforms. Together, these factors enable the deployment of general-purpose robotic capabilities in human-built spaces without extensive retrofitting. Humanoids share a partial overlap in capabilities with humans: they can perform certain tasks humans can, while humans retain abilities that humanoids do not, and vice versa. This asymmetric overlap gives rise to distinct strengths, limitations, and failure modes that must be accounted for in design and evaluation (Gundawar et al., 2025).
2. **From machines as tools to social actors.** Unlike conventional machines that are interpreted primarily as tools, humanoids' human-like appearance and motion cues lead people to attribute intent, understanding, and social competence to them. As a result, users naturally expect humanoids to communicate, coordinate, and respond in human-like ways, even when their underlying capabilities do not fully support such expectations (Mori et al.,

2012). This expectation shift affects how instructions are given, how errors are interpreted, and how trust is formed during interaction. When humanoid behavior falls short of these socially inferred expectations, it can lead to confusion, misuse, or erosion of trust, highlighting the need to explicitly account for expectation management, transparency, and communicative behavior in humanoid system design and evaluation.

3. **From technical safety to socio-ethical governance.** Traditional automation emphasizes technical safety guarantees such as collision avoidance, fault tolerance, and reliability under predefined conditions (Brunke et al., 2022; Corso & Kochenderfer, 2020; Lasota et al., 2017). In contrast, humanoids introduce broader trust, safety, and ethical considerations that arise from their anthropomorphic form, autonomy, and intelligence (Winfield & Jirotka, 2018). Because humanoids operate in shared social spaces and are perceived as agents rather than tools, questions of accountability, appropriate behavior, and responsibility become inseparable from system performance. This shift necessitates governance frameworks that extend beyond engineering safeguards to include social norms, ethical constraints, and institutional oversight, addressing not only whether a humanoid can act safely, but whether it should act, under whose authority, and with what consequences when failures occur.

Taken together, these shifts indicate that the promise of humanoids is not defined solely by capability, such as mobility or dexterity, but by compatibility with human environments, norms, and expectations. Effective deployment therefore requires legible behavior, reliable communication, and sustained co-adaptation to human practices and social conventions (Hancock et al., 2011; Sanneman & Shah, 2020). Achieving this compatibility demands that embodiment, cognition, social interaction, and ethics be treated as co-equal design constraints rather than isolated considerations.

As summarized in Figure 1, we therefore structure the Humanoid Factors framework around four dimensions—**Physical (P)**, **Cognitive (C)**, **Social (S)**, and **Ethical (E)**—that capture the key ways humanoid systems differ from both traditional automation and humans themselves. Each dimension addresses distinct design challenges and evaluation criteria that emerge when intelligent, human-shaped agents operate in shared human spaces, and together they provide a principled foundation for guiding humanoid design and assessing human–humanoid coexistence.

The complete Humanoid Factors framework (`https://arxiv.org/abs/2602.10069`) decomposes each pillar into multiple layers. Here we focus on one layer of the Cognitive pillar: using explicit, validated human models as evaluation primitives, motivated by the goal of introducing explicit human models into the reasoning process.

2.1 Cognitive (**C**) compatibility as a design and evaluation target

The Cognitive Pillar specifies how a humanoid transforms sensing into understanding, intention, and action over time. Rather than prescribing specific algorithms or architectures, we organize cognition into a layered reasoning model that distinguishes (i) how the robot interprets what is *happening*, (ii) how it decides what to *do next*, (iii) how it *self-regulates* its behavior to remain safe, legible, and intelligible, and (iv) how it manages memory and cognitive load to *sustain* effective operation over time (Sanneman et al., 2023). This question-driven structure clarifies what sensor measurements mean for the robot's internal state and behavior—progressing from situational awareness (Endsley, 2021; Endsley et al., 2000; Sanneman & Shah, 2022) to planning, guardrails, and memory—while treating computational limits, latency, training coverage, and operational design domain (ODD) constraints as first-order design requirements rather than afterthoughts.

In humanoids, cognition is implemented through AI systems, and this pillar organizes the capabilities those systems must support—whether realized explicitly through traditional modular AI components or implicitly through modern monolithic AI foundation models—into a coherent layered structure.

2.1.1 Layer C2: Reasoning, Planning, & Execution (What to *do next*)

Once a humanoid has established situational awareness, it must determine how to act. This layer concerns the cognitive processes by which a humanoid selects goals, reasons about alternatives, and translates intent into executable behavior under uncertainty, time pressure, and resource con-

straints. Rather than prescribing a single control paradigm, this layer characterizes the functional roles required to support human-like decision-making, including goal prioritization, planning under uncertainty, and adaptive execution.

Reasoning, planning, and control separation. To support both flexibility and auditability, modern humanoid systems benefit from a conceptual separation between three cognitive functions:

- **Reasoning** evaluates uncertainty (Senanayake, 2025), assesses whether sufficient information is available to act, and considers alternative courses of action (e.g., "Should I proceed, slow down, or ask for clarification?"). This role corresponds to meta-cognitive (Dunlosky & Metcalfe, 2008) decision checks studied in human factors, such as confidence assessment and risk sensitivity (Endsley, 2017).
- **Planning** translates high-level intent into structured subgoals and selects appropriate skills to achieve them. Planning must handle underspecified instructions, competing objectives, and dynamic environments, while maintaining internally consistent goal hierarchies. Generating intermediate goals and rationales supports downstream legibility and verification.
- **Execution** realizes planned actions through low-latency motor control and skill execution, adapting online to disturbances and human motion. Execution prioritizes responsiveness and safety while remaining consistent with the planner's intent.

This separation mirrors well-established distinctions in human cognition between deliberation, intention formation, and motor execution, and provides clear interfaces for evaluation and failure diagnosis. Where helpful for intuition and cross-disciplinary grounding, map these functions to brain systems: higher-level planning and goal maintenance are associated with prefrontal and premotor networks, online sensorimotor transformation and execution with motor cortex and cerebellar circuits, and episodic/relational memory with hippocampal systems (Henschke & Pakan, 2023). These analogies are heuristic (robots are not brains), but they are useful for deriving evaluation primitives—e.g., short-term working memory capacity, predictive timing accuracy, and consolidation of episodic experience.

Decision-making under uncertainty and time constraints. Human environments rarely permit fully informed or perfectly timed decisions. Accordingly, this layer emphasizes decision-making under uncertainty (Kochenderfer, 2015), including explicit trade-offs between speed and accuracy, confidence thresholds for action (Murthy & Sanneman, 2025), and graceful degradation when information, compute, or energy budgets are limited. Planning horizons and execution timing should reflect human-comfortable tempos, supporting coordination and reducing surprise during joint action. For instance, as we will *demonstrate in Section 3, classical models such as Fitts' Law* (MacKenzie, 1992), long used in human–computer interface design, remain informative for setting timing targets in pointing, handover, and placement tasks and for defining acceptable speed–accuracy envelopes.

Goal management and adaptability. This layer also governs how goals are maintained, revised, or abandoned over time. Humanoids must reconcile long-term objectives with short-term contingencies, resolve conflicts between safety, efficiency, and user intent, and adapt plans in response to unexpected events. These capabilities align with human factors research on goal switching, workload management, and adaptive behavior in complex systems.

3 Case study: Fitts' Law as an evaluation primitive for humanoid control

We now ground the Humanoid Factors perspective with a concrete case study on a humanoid platform. While comprehensive evaluation should ultimately account for all four pillars, here we focus on the Cognitive–Physical interface and ask a targeted question: *Does training a robot to achieve geometric task success also yield cognitively legible motion patterns that support predictable and comfortable human interaction?*

3.1 HUMAN-CENTERED MOTIVATION

Humanoids are expected to operate across a wide range of settings. While some deployments may occur in relatively isolated environments such as factories, many envisioned applications involve close, continuous interaction with people—for example in homes, clinics, or caregiving contexts. In such settings, humanoids must engage in everyday collaborative actions, such as *handing objects to a person*, in ways that align with human expectations formed through human–human interaction. In these situations, humans will increasingly encounter humanoids built by different manufacturers and powered by diverse AI models. From the human's perspective, however, these distinctions are largely irrelevant: regardless of the underlying model or brand, the humanoid is expected to behave in ways that are predictable, understandable, and safe. Making sure this expectation is met is therefore not just a deployment concern, but a responsibility of the AI models that generate the humanoid's behavior.

Current evaluation practices for humanoid AI models, however, remain predominantly *task completion-centric*, emphasizing whether a task is completed or a predefined checkpoint is reached. From a HoF perspective, this is insufficient. In human-facing settings, *how a humanoid moves is as critical as whether it completes the task "somehow."* For a humanoid to coexist comfortably with people, its motion must be *legible*: a human observer should be able to intuitively anticipate the robot's goal, timing, and level of commitment from its movement alone (Dragan et al., 2013).

Human sensitivity to motion predictability is not arbitrary; it is shaped by robust regularities in human motor behavior that govern how speed, timing, and variability adapt to task demands (often modeled by psychophysical laws). When a humanoid completes a task while violating these implicit expectations—for instance by moving with uniform speed across easy and difficult segments, or by hesitating without clear intent—it can create cognitive dissonance for the human partner, undermining trust, comfort, and perceived safety (Dragan et al., 2013; Howell et al., 2023).

3.2 FITTS' LAW

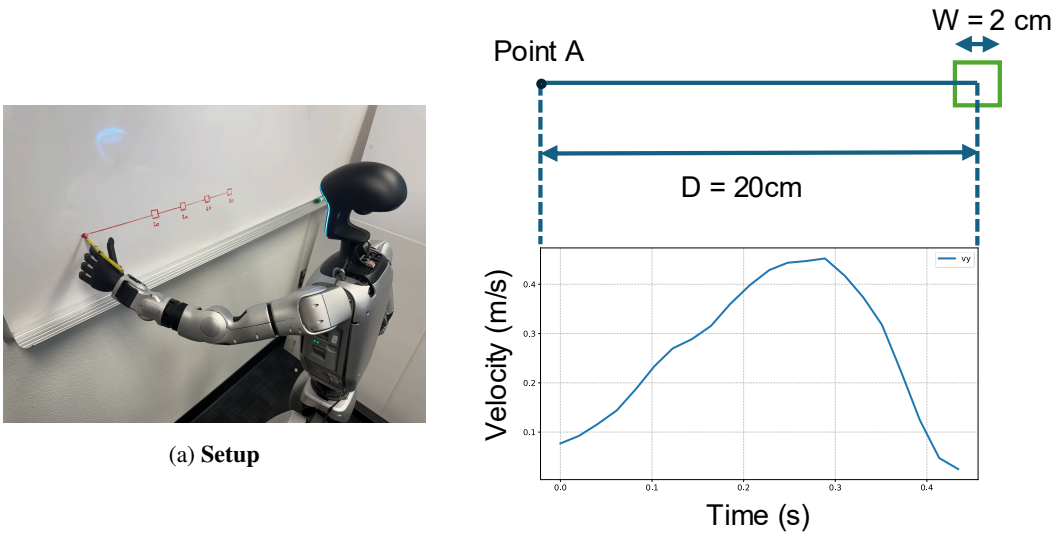
One of the most widely studied psychophysical models of human movement is Fitts' Law (Fitts, 1954; MacKenzie, 2018), which describes how the time required to complete a goal-directed movement scales with task difficulty. Originally developed to model pointing and reaching actions, it has since been validated across a broad range of real-world motor activities, including computer input device usage analysis (Senanayake et al., 2013; Senanayake & Goonetilleke, 2013), tool use (Silva et al., 2016), surgical and rehabilitation tasks (McCrea & Eng, 2005; Zimmerli et al., 2012), vehicle and cockpit controls (Large et al., 2015; Xie et al., 2023), and everyday object manipulation (Kantowitz & Elvers, 1988; Thumser et al., 2018). Across these domains, the law captures how humans systematically adapt movement speed and timing in response to precision demands, producing characteristic motion profiles that are highly predictable to observers. For instance, in human–human object handover, one subcase of the object transportation phase (Kopnarski et al., 2023) involves moving an object toward a fixed target location corresponding to the partner's hand, yielding timing and velocity patterns that resemble those of reaching actions.

In psychophysics, the time required to complete such a reaching or pointing movement (*Movement Time*, MT) is commonly modeled as a linear function of an *Index of Difficulty* (ID_F) that depends on movement distance D and target width W:

$$MT = a + b\, ID_F, \qquad ID_F = \log_2\left(\frac{2D}{W}\right).$$

We use this common formulation of Fitts' Law throughout; in our experiments W is fixed and we vary D to sweep task difficulty.

Traditionally, from an engineering perspective, AI-based robot controllers or deterministic planners typically optimize for trajectory efficiency, torque, or smoothness (e.g., minimum jerk) without inherently adhering to Fitts-like scaling (Gasparetto et al., 2015; Li et al., 2023) because robots are not constrained by biological neuromotor limits (Takeda et al., 2019). Consequently, it is possible to use Fitts' Law not as a control limit, but as a benchmark for measuring the "naturalness" and usability of assisted teleoperation and mobile manipulation (Pan et al., 2024; Wan et al., 2023). Similarly, Fitts' Law can be used to test behavioral cloning (BC)-based neural networks (Torabi et al., 2018) or any humanoid foundation model.



(a) **Setup**

Figure 2: **Experimental Environment and Data.** (a) Unitree G1 robot setup. (b) Example end-effector velocity profile illustrating the pencil tip motion from the start point to the target region for a single demonstration ($D = 20$ cm and $W = 2$ cm).

Using HoF for Evaluation:
1. Train an AI model (e.g., Behavioral cloning).
2. Select a relevant HoF layer (C2: Reasoning, Planning, & Execution).
3. Identify validation models associated with that layer (e.g., Fitts' Law).
4. Test statistical alignment with validation models beyond task-centric metrics.

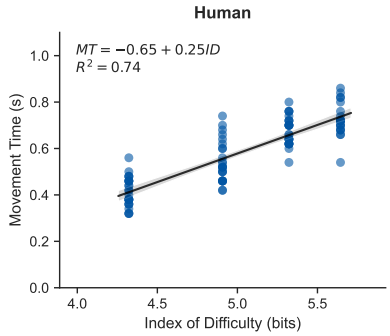
3.3 Experiment

Hypothesis. We hypothesize that although a BC policy (with only proprioception data) can achieve task completion, its execution time will not follow the Fitts' Law relationship between movement time MT and index of difficulty ID_F. In particular, the fitted MT–ID_F trend for the learned policy will deviate from that of human demonstrations (e.g., reduced correlation and/or different slope), reflecting a failure to reproduce human temporal scaling with difficulty.

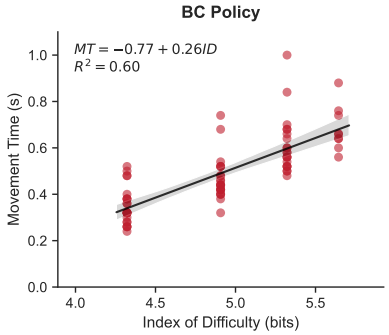
Hardware platform. We conducted experiments using the Unitree G1 humanoid robot, as shown in Figure 2. The experiment focused on a pointing task using the left arm (7-DoF). A pencil was attached to the robot's hand to serve as a precise end-effector extension for the pointing task. We only controlled four joints (left shoulder yaw, pitch, roll; left elbow) while locking the wrist and all other joints. We only command a high-level sequence of target joint angles along the trajectory suggested by the AI model. Low-level motor control is handled by the robot's built-in PID controllers. We verified that the same PID stack can accurately replay recorded demonstration trajectories, suggesting that the Fitts' Law breakdown during learned rollouts is unlikely to arise from PID tracking alone.

Task. The robot was initialized at its neutral starting position. The goal was to move the end-effector into a square target region ($W = 0.02$ m). We varied the distance (D) across four levels: 0.20 m, 0.30 m, 0.40 m, and 0.50 m. A human operator provided demonstrations via kinesthetic teaching, physically guiding the robot arm (the pencil tip) from Point A to the target box as quickly as possible, resulting in 99 demonstrations after discarding recording errors or target misses.

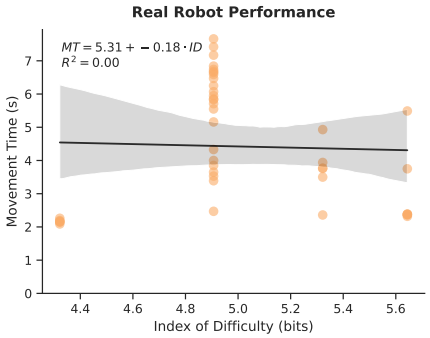
AI Model Training and Deployment. We consider the task of training a robot to imitate human behavior from multiple demonstrations of the task described above. We trained a conditional Behavior Cloning (BC) policy with multi-layer perceptron (i.e., fully connected) neural network by



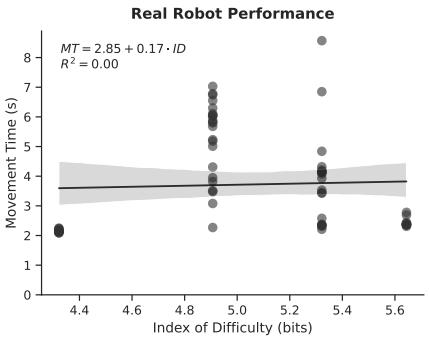
(a) **Human Baseline:** Strong linear fit ($R^2 = 0.741$), indicating natural speed-accuracy trade-off.



(b) **Simulation Rollout:** Parallel slope to human data ($R^2 = 0.596$), suggesting successful velocity cloning.



(c) **Real World ($K_p = 60$):** Significant degradation. The low stiffness appears insufficient to maintain the commanded pose under load, leading to drift and highly variable movement times.



(d) **Real World ($K_p = 80$):** Higher stiffness improves success rate but does not restore Fitts' Law linearity. Performance still degrades at the largest reach distance ($D = 0.5\,m$), consistent with increased gravity loading.

Figure 3: **Does the robot follow Fitts' Law?** (a) The human operator establishes a cognitive baseline. (b) The simulation captures the general trend but introduces variance. (c-d) Real-world physical constraints (e.g., gravity, friction) can decouple the motion from the cognitive plan, with higher stiffness ($K_p = 80$) offering only marginal recovery of linearity. This mismatch behavior poses a risk to human anticipation and trust.

minimizing the mean squared error (MSE) loss between predicted joint angles and target joint angles. After training, the model was evaluated first in simulation using MuJoCo and then on the real robot. We used an open-loop autoregressive strategy in which the model's previously predicted joint angles were fed as input for the next step, rather than using the robot's actual sensor readings. This isolates and tests the stability of the model's internal plan independent of corrective feedback.

3.4 Results and Discussions

We analyzed the adherence to Fitts' Law across the human baseline, simulation, and real-world deployment.

Human baseline (gold standard). The kinesthetic demonstrations (see Figure 3a) exhibited a strong adherence to Fitts' Law ($R^2 = 0.741$). This confirms the validity of our setup: even when manipulating a robot arm, the human operator naturally engaged in a speed-accuracy trade-off.
Simulation rollout (cognitive success). As shown in Figure 3b, the policy generated trajectories that roughly paralleled the human baseline in the MuJoCo physics simulation. The fit was moderate ($R^2 = 0.596$), while the slope of the robot's regression line was nearly parallel to the human's. This implies the BC model successfully cloned the planning strategy (velocity profile) of the human.
Real-world deployment (physical failure). We deployed the policy on the physical Unitree G1

robot at two stiffness levels. For low stiffness ($K_p = 60$), as seen in Figure 3c, the Fitts' Law relationship collapsed. The arm was too compliant to hold the target pose against gravity, leading to excessive drift and erratic movement times. For high stiffness ($K_p = 80$), increasing the gain (Figure 3d) improved the success rate by rigidly enforcing the trajectory. However, it did not restore the linear trend. The robot still struggled with the "Lever Arm" effect at large distances ($D = 0.5m$), proving that purely geometric gains cannot compensate for a lack of internal gravity modeling.

The disparity between the simulation success and the physical failure reveals two critical blind spots in standard AI training, validating our framework.

The "orbiting" phenomenon (Cognitive Pillar). In simulation and reality, the robot sometimes arrived at the target area but failed to stop, instead "orbiting" or oscillating around the $2cm$ target. We attribute this to the MSE loss used in training. MSE penalizes large errors heavily but provides small gradients for small errors. The robot learns to *go* toward the target but never learns to *stop*. This violates the Cognitive Pillar: a human partner cannot interact with a robot that hovers indecisively.

Pose-dependent gravity load (Physical Pillar). The most severe failures occurred at the extremes of the reach, highlighting distinct physical violations. In the far field ($D = 0.5m$), the failure is driven by the *Lever Arm Effect*. At maximum extension, the gravity torque on the shoulder (Joint 0) is maximized. Since our open-loop deployment is blind to torque loading, the arm physically sags below the target despite the correct kinematic plan. Conversely, in the near field ($D = 20$cm), failure is attributed to a *low-manipulability* folded configuration (i.e., proximity to a kinematic singularity where the Jacobian becomes ill-conditioned). In this regime, small errors in the joint-space policy can produce disproportionately large and unstable end-effector deviations, exacerbating the "orbiting" behavior and drastically increasing Movement Time.

Incorporating human cognitive principles into humanoid foundation models. A BC neural network trained only on kinematics (joint angles) lacks an internal model of compliance and dynamics. To satisfy Physical-Cognitive Pillars of HoF, AI training and tuning must account for real-world software–hardware interactions, ensuring that learned policies reflect not only kinematic feasibility but also the physical realities of compliance, dynamics, gravity, and singular configurations.

Modern LLMs are typically trained using a three-stage recipe consisting of **pre-training, mid-training, and post-training**. Pre-training provides broad statistical competence by learning general representations from large-scale, heterogeneous data, but remains largely agnostic to embodiment, task structure, and human cognitive constraints (Tu et al., 2025). In the humanoid setting, this stage offers a natural entry point for embedding generic cognitive priors, such as multimodal situational awareness, temporal abstraction, and coarse world models that support perception–action coupling. Mid-training adapts the foundation model to specific domains, embodiments, and operational contexts, making it the primary stage for injecting humanoid-specific cognitive structure, including goal hierarchies, task decomposition, affordance reasoning, and planning biases aligned with human environments (Tu et al., 2025; Mo et al., 2025; Liu et al., 2025). At this stage, cognition can be shaped through structured curricula, simulator-grounded interaction data, and embodiment-aware objectives that reflect human-scale constraints and expectations. Post-training aligns the model with human values, preferences, and safety norms using human feedback, preference optimization, and deployment-aware fine-tuning (Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). This stage is particularly well suited for embedding human cognitive norms, such as uncertainty awareness, self-regulation, legibility, and calibrated intervention, ensuring that humanoid behavior remains predictable and interpretable in human-shared spaces. We discuss these aspects at length in the Humanoid Factors framework (`https://arxiv.org/abs/2602.10069`).

4 Conclusion

Foundation model humanoids increase the urgency of measuring not only *what* robots achieve, but *how* they behave in ways humans can predict and are comfortable. We showed that a compact cognitive-science model (Fitts' Law) can serve as a practical evaluation primitive that exposes failure modes missed by completion-centric robotics metrics. We believe cognitive primitives similar to these can provide a tractable bridge between human cognitive science and foundation models in embodied systems.

REFERENCES

Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. Progress and prospects of the human–robot collaboration. *Autonomous robots*, 42(5):957–975, 2018.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL `https://crfm.stanford.edu/assets/report.pdf`.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

Longbing Cao. Humanoid robots and humanoid ai: Review, perspectives and directions. *ACM Comput. Surv.*, 58(4), October 2025. ISSN 0360-0300. doi: 10.1145/3770574. URL `https://doi.org/10.1145/3770574`.

Anthony Corso and Mykel J Kochenderfer. Interpretable safety validation for autonomous vehicles. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6. IEEE, 2020.

Kerstin Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480):679–704, 2007.

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. IEEE, 2013.

John Dunlosky and Janet Metcalfe. *Metacognition*. Sage Publications, 2008.

Mica R Endsley. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*, pp. 9–42. Routledge, 2017.

Mica R Endsley. Situation awareness. *Handbook of human factors and ergonomics*, pp. 434–455, 2021.

Mica R Endsley, Daniel J Garland, et al. Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 1(1):3–21, 2000.

Paul M Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

Alessandro Gasparetto, Paolo Boscariol, Albano Lanzutti, and Renato Vidoni. *Path Planning and Trajectory Planning Algorithms: A General Overview*, pp. 3–27. Springer International Publishing, 2015. ISBN 978-3-319-14705-5. doi: 10.1007/978-3-319-14705-5_1.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the Physical World. *arXiv preprint arXiv:2503.20020*, 2025.

Atharva Gundawar, Som Sagar, and Ransalu Senanayake. PAC bench: Do foundation models understand prerequisites for executing manipulation policies? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=bbQV3GQ6Zy.

Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

Julia U Henschke and Janelle MP Pakan. Engaging distributed cortical and cerebellar networks through motor execution, observation, and imagery. *Frontiers in systems neuroscience*, 17: 1165307, 2023.

Pierce Howell, Jack Kolb, Yifan Liu, and Harish Ravichandar. The effects of robot motion on comfort dynamics of novice users in close-proximity human-robot interaction. *arXiv preprint arXiv:2308.01466*, 2023.

Barry H Kantowitz and Greg C Elvers. Fitts' law with an isometric controller: effects of order of control and controldisplay gain. *Journal of Motor Behavior*, 20(1):53–66, 1988.

Laurence Kaufmann and Fabrice Clément. How culture comes to mind: From social affordances to cultural analogies. *Intellectica*, 46:221–250, 2007.

Charles C. Kemp, Aaron Edsinger, and Eduardo Torres-Jara. Challenges for robot manipulation in human environments [grand challenges of robotics]. *IEEE Robotics Automation Magazine*, 14 (1):20–29, 2007. doi: 10.1109/MRA.2007.339604.

Mykel J Kochenderfer. *Decision making under uncertainty: theory and application*. MIT press, 2015.

Lena Kopnarski, Julian Rudisch, and Claudia Voelcker-Rehage. A systematic review of handover actions in human dyads. *Frontiers in Psychology*, 14:1147296, 2023.

Dávid Kóczi and József Sárosi. Safety engineering for humanoid robots in everyday life—scoping review. *Electronics*, 14(23), 2025. ISSN 2079-9292. URL https://www.mdpi.com/2079-9292/14/23/4734.

David R Large, Elizabeth Crundall, Gary Burnett, and Lee Skrypchuk. Predicting the visual demand of finger-touch pointing tasks in a driving context. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 221–224, 2015.

Przemyslaw A Lasota, Terrence Fong, Julie A Shah, et al. A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics*, 5(4):261–349, 2017.

S. Li, W. Qi, Y. Hu, H. R. Karimi, and G. Ferrigno. Human-like motion planning of robotic arms based on human arm motion patterns. *Robotica*, 41(1):259–276, 2023.

Zhengzhong Liu, Liping Tang, Linghao Jin, Haonan Li, Nikhil Ranjan, Desai Fan, Shaurya Rohatgi, Richard Fan, Omkar Pangarkar, Huijuan Wang, et al. K2-v2: A 360-open, reasoning-enhanced llm. *arXiv preprint arXiv:2512.06201*, 2025.

I Scott MacKenzie. Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction*, 7(1):91–139, 1992.

I Scott MacKenzie. Fitts' law. *The wiley handbook of human computer interaction*, 1:347–370, 2018.

Patrick H McCrea and Janice J Eng. Consequences of increased neuromotor noise for reaching movements in persons with stroke. *Experimental brain research*, 162(1):70–77, 2005.

Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. Mid-training of large language models: A survey. *arXiv preprint arXiv:2510.06826*, 2025.

Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012.

Anil B Murthy and Lindsay Sanneman. Llms need to go beyond computational confidence metrics to establish trust. In *Proceedings of the AAAI Symposium Series*, volume 7, pp. 131–136, 2025.

Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *2013 8th ACM/IEEE international conference on human-robot interaction (HRI)*, pp. 33–40. IEEE, 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Jiahe Pan, Jonathan Eden, Denny Oetomo, and Wafa Johal. Using fitts' law to benchmark assisted human-robot performance. *arXiv preprint arXiv:2412.05412*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Diego Rodriguez-Guerra, Gorka Sorrosal, Itziar Cabanes, and Carlos Calleja. Human-robot interaction review: Challenges and solutions for modern industrial environments. *Ieee Access*, 9: 108557–108578, 2021.

Lindsay Sanneman and Julie A Shah. Trust considerations for explainable robots: A human factors perspective. *arXiv preprint arXiv:2005.05940*, 2020.

Lindsay Sanneman and Julie A Shah. The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems. *International Journal of Human–Computer Interaction*, 38(18-20):1772–1788, 2022.

Lindsay Sanneman, Mycal Tucker, and Julie Shah. An information bottleneck characterization of the understanding-workload tradeoff. *arXiv preprint arXiv:2310.07802*, 2023.

Ransalu Senanayake. *The Role of Predictive Uncertainty and Diversity in Embodied AI and Robot Learning*, pp. 148–182. Cambridge University Press, 2025.

Ransalu Senanayake and Ravindra S Goonetilleke. Superiority of freehand pointing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pp. 1639–1642. SAGE Publications Sage CA: Los Angeles, CA, 2013.

Ransalu Senanayake, Errol R Hoffmann, and Ravindra S Goonetilleke. A model for combined targeting and tracking tasks in computer applications. *Experimental brain research*, 231(3):367–379, 2013.

Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

Paula L Silva, Reinoud J Bootsma, Priscilla Rezende Pereira Figueiredo, Bruna Silva Avelar, André Gustavo Pereira de Andrade, Sérgio T Fonseca, and Marisa Cotta Mancini. Task difficulty and inertial properties of hand-held tools: An assessment of their concurrent effects on precision aiming. *Human movement science*, 48:161–170, 2016.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Leila Takayama, Wendy Ju, and Clifford Nass. Beyond dirty, dangerous and dull: what everyday people think robots should do. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 25–32, 2008.

Misaki Takeda, Takanori Sato, Hisashi Saito, Hiroshi Iwasaki, Isao Nambu, and Yasuhiro Wada. Explanation of fitts' law in reaching movement based on human arm dynamics. *Scientific reports*, 9(1):19804, 2019.

Zachary C Thumser, Andrew B Slifkin, Dylan T Beckler, and Paul D Marasco. Fitts' law in the control of isometric grip force with naturalistic targets. *Frontiers in psychology*, 9:560, 2018.

Yuchuang Tong, Haotian Liu, and Zhengtao Zhang. Advancements in humanoid robots: A comprehensive review and future prospects. *IEEE/CAA Journal of Automatica Sinica*, 11(2):301–328, 2024.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 4950–4957. AAAI Press, 2018. ISBN 9780999241127.

Chengying Tu, Xuemiao Zhang, Rongxiang Weng, Rumei Li, Chen Zhang, Yang Bai, Hongfei Yan, Jingang Wang, and Xunliang Cai. A survey on llm mid-training. *arXiv preprint arXiv:2510.23081*, 2025.

Yuhui Wan, Jingcheng Sun, Christopher Peers, Joseph Humphreys, Dimitrios Kanoulas, and Chengxu Zhou. Performance and usability evaluation scheme for mobile manipulator teleoperation. *IEEE Transactions on Human-Machine Systems*, 53(5):844–854, 2023.

Alan FT Winfield and Marina Jirotka. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180085, 2018.

Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Shuo Jiang, Bin He, and Qian Cheng. Robot learning in the era of foundation models: A survey. *Neurocomputing*, pp. 129963, 2025.

Yubin Xie, Ronggang Zhou, and Jianhong Qu. Fitts' law on the flight deck: evaluating touchscreens for aircraft tasks in actual flight scenarios. *Ergonomics*, 66(4):506–523, 2023.

Wei Xu and Zaifeng Gao. Enabling human-centered ai: A methodological perspective. *arXiv preprint arXiv:2311.06703*, 2023.

Lukas Zimmerli, Carmen Krewer, Roger Gassert, Friedemann Müller, Robert Riener, and Lars Lünenburger. Validation of a mechanism to balance exercise difficulty in robot-assisted upper-extremity rehabilitation after stroke. *Journal of neuroengineering and rehabilitation*, 9(1):6, 2012.