When Stopping Is Falling: A Safe-State Reachability Framework for Protective-Stop Design in Dynamically Stable Industrial Robots

Jherrod Thomas
Independent Researcher — The Lion of Functional Safety
jherrodthomas.com · October 2026 · Field Notes Vol. 1, No. 12

***Abstract*—Every protective-stop convention in machinery safety assumes that removing power drives a machine toward its safe state. For a dynamically stable robot (a legged humanoid that needs active control to remain upright) the assumption fails: de-energization converts a controlled machine into an unguided falling mass, and the only safe posture that survives power removal must itself be reached under power, with working actuators and a working balance controller. The safe state therefore has a probability of not being reached, and that probability is multiplied by the rate at which the protective function is demanded. We formalize this as a safe-state reachability framework comprising a four-rung safe-state ladder, a fall-zone radius, a hazardous-event rate that separates demand-triggered reach failures from spontaneous balance loss, a reach-failure decomposition with pre-demand diagnostic coverage, a break-even condition between balanced standstill and seated descent, and the demonstration burden each target implies. Applied to the published parameters of a 129 kg, 1.81 m bipedal humanoid launched for barrier-free industrial work in September 2026, the framework shows that a seat-on-approach policy exceeds the Performance Level d budget by four orders of magnitude at a realistic approach rate, that balanced standstill is the safer proximity response unless the seat reach-failure probability falls below about 4×10^{-8}, and that the seated descent is properly a de-energization procedure, not a proximity response. Seven requirements are derived for the forthcoming ISO 25785-1 regime.**

***Index Terms*—ISO/CD 25785-1, ISO 13849-1, IEC 60204-1 Stop Categories, ISO 13850, Dynamically Stable Robots, Humanoid Robots, Safe State, Fall Mitigation, Speed and Separation Monitoring.**

I. Introduction

Machinery safety has a reflex older than any of its current standards: when in doubt, take the power away. An emergency-stop mushroom, a safety relay dropping out when a light curtain is broken, a drive executing safe torque off — each assumes that a machine with no energy at its actuators is a machine that is no longer dangerous. IEC 60204-1 codifies the reflex as Stop Category 0, immediate removal of power to the machine actuators, and makes it the reference against which the controlled stops of Categories 1 and 2 are defined [1]. ISO 13850 permits only Categories 0 and 1 for the emergency-stop function [2]. ISO 13849-1 then supplies the reliability vocabulary — Performance Level, probability of dangerous failure per hour, diagnostic coverage — for the control channel that commands the stop [3]. The entire edifice rests on the premise that the de-energized state is reachable without effort: it is what the machine falls into when everything else fails.

For a dynamically stable robot, the machine literally falls into it. ISO/CD 25785-1, the first Type C standard written for this class, defines its scope by exactly this property: a robot with actively controlled stability, "a robot that requires an active control in order to remain balanced and could become unstable in the absence of power" [4]. For a bipedal humanoid, removing power converts a controlled machine into an inverted pendulum whose only stable equilibrium is on the floor. The energy involved is not marginal. A 129 kg robot with a center of mass one meter above the ground releases on the order of 1.3 kJ in a full fall, roughly three orders of magnitude above the transferable-energy values that collaborative-robot practice treats as tolerable for transient contact [5], [6].

The question is no longer academic. On September 15, 2026, Agility Robotics launched Digit 5, a 1.81 m, 129 kg bipedal humanoid with a 2.2 m reach and a 22.7 kg repeat-lift payload, described as engineered to work in proximity to people without protective barriers [7]. The published safety architecture separates proprietary AI-based person detection from an independent safety controller, and lists three responses to a person within an unsafe distance: avoiding the person, stopping, or assuming a seated position [7]. The standard that would govern that architecture, ISO/CD 25785-1, closed its committee-draft comment period on July 8, 2026 and remains at stage 30.60 [4]. The IEEE Humanoid Study Group's 2025 report on future standards identified the same gap: humanoids collapse when power is cut, and the existing mobile-robot standards assume they do not [8].

The industrial injury record explains why the gap matters. Sanders, Şener and Chen identified 77 robot-related accidents in OSHA Severe Injury Reports between 2015 and 2022, 23 of them involving mobile robots, whose injuries were predominantly fractures to the legs and feet [9]. Those mobile robots were statically stable. OSHA still has no robotics-specific standard and addresses robot injuries through machine guarding, lockout/tagout and the General Duty Clause [10]. Barrier-free legged robots add a hazard none of those injuries involved: a fall from height, by the robot, onto the person.

This paper addresses a single research question. *How should the protective-stop function of a dynamically stable robot be specified and verified when its safe state must be reached under*

power and can therefore fail to be reached? Our answer is that the probability of failing to reach the safe state is a per-demand quantity, which makes the demand rate a safety parameter rather than an operational one, and that the correct response to a person depends on which of two hazard terms dominates. The contributions are:

- A **safe-state ladder** of four states (balanced standstill, seated or braced posture, controlled fall, and the uncontrolled fall they exist to prevent) characterized by their energy dependence during the transition and while held, with a mapping onto the IEC 60204-1 stop categories that shows the seated descent is a Category 1 stop rather than a deviation from one.
- **Eight numbered relations**: a balance-loss response window, a fall-zone radius, a hazardous-event rate separating demand-triggered reach failures from spontaneous balance loss, a reach-failure decomposition with pre-demand diagnostic coverage, a break-even condition between standstill and seated descent, a zero-failure demonstration bound, a hold-up energy condition, and a demand budget.
- A **structural result**: under a seat-on-approach policy, the person-approach rate multiplies the reach-failure probability, so the protective function's own demand rate becomes the dominant hazard term, inverting the usual intuition that more frequent protective stops mean a safer machine.
- A **worked application** to the published Digit 5 parameters with explicitly flagged assumptions, and seven derived requirements anchored to ISO 13849-1, IEC 60204-1, ISO 13850, ISO 13855 and the draft ISO 25785-1.

Section II reviews background and locates the gap. Section III develops the framework. Section IV applies it. Section V discusses limitations and where the standards lens stops. Section VI concludes.

II. Background and Related Work

A. Stop Categories and the Fixed-Base Assumption

IEC 60204-1 defines three stop categories. Category 0 removes power immediately; Category 1 is a controlled stop with power available to the actuators to achieve the stop, followed by removal of power when the stop is achieved; Category 2 is a controlled stop with power left available [1]. ISO 10218-1:2025 adopts these for industrial robots and, in its third edition, replaces the former safety-rated monitored stop with monitored standstill, a Category 2 condition supervised by a safety function [6]. ISO 3691-4 and the ANSI/A3 R15.08 series extend the logic to driverless trucks and industrial mobile robots, with protective fields whose size is a function of speed and stopping distance [11]. Every one of these documents assumes static stability: whatever stop category is selected, the robot remains upright when it is achieved and remains upright when power is subsequently removed. Under that assumption, Category 0 trades a longer stopping distance for an unconditionally reachable end state, and Category 1 trades a dependency on power for a shorter one. Neither trade includes the possibility that the end state itself is unreachable.

B. Balance, Capture and Fall Prediction

The biped-control literature supplies the physics the safety standards lack. The zero-moment point formalized the condition under which a legged robot's support polygon can carry its dynamic load [12]. The linear inverted pendulum reduced balance to a tractable model whose time constant $\tau = \sqrt{L/g}$ sets the speed at which an unbalanced robot diverges [13]. Pratt and colleagues introduced the capture point and capture region — the ground locations at which a robot must step to come to a complete stop — and with them a principled answer to when balance is still recoverable [14]. Kalyanakrishnan and Goswami cast fall prediction as a learned classifier that switches the robot from balance maintenance to fall control with sufficient lead time while minimizing false alarms [15]. That last framing matters for safety engineering because it makes fall prediction a detection function with a false-negative and a false-positive rate, which is precisely the form ISO 13849-1 cannot accommodate when the detector is learned rather than designed [3].

C. Fall Mitigation

Once a fall is unavoidable, its consequences can still be shaped. Goswami *et al.* demonstrated direction-changing fall control through stepping and inertia shaping, explicitly motivated by steering the fall away from a person or delicate object [16]. Kumar, Ha and Liu learned a unified policy that minimizes maximal impulse during falls [17]. Subburaman *et al.* surveyed fall prediction, controlled fall and fall recovery and identified the lack of standardized metrics as an open problem [18]. Recent work moves to learned whole-body protective control: SafeFall trains a policy for protective falling on a full-size humanoid [19], and Xu *et al.* learn a unified fall-safety policy from few demonstrations [20]. This literature is concerned with damage to the robot and, secondarily, to the environment; none of it states its performance as a coverage factor that a safety case could credit. The DARPA Robotics Challenge finals in 2015 remain the largest public record of legged-robot falls under task load, with multiple teams' robots falling during the competition runs [21].

D. Human-Robot Safety and Validation Burden

The human-robot interaction literature provides the separation and contact models. Haddadin, Albu-Schäffer and Hirzinger measured injury mechanisms in robot impacts and established mass and velocity as the governing variables [5]. Marvel and Norcross analyzed the speed-and-separation monitoring protective distance in detail and its sensitivity to latency and measurement uncertainty [22]. Surveys by Villani *et al.*, Lasota *et al.* and Guiochet *et al.* catalogue the mechanisms and the verification difficulty of advanced robots [23], [24], [25]. From the automated-driving domain, Kalra and Paddock and Koopman and Wagner established that demonstrating rare-event reliability by testing alone requires exposure that is often infeasible [26], [27]; the same arithmetic applies to the reach-failure probability developed here.

E. The Gap

No existing work connects these strands. The stop-category standards assume a reachable safe state; the fall literature optimizes the fall without assigning it a reliability claim; and the separation literature sizes protective fields for robots that stop where they stand. What is missing is a way to write the probability that a dynamically stable robot fails to reach its safe state into the same ledger as the Performance Level of the function that commanded it.

III. Safe-State Reachability Framework

A. The Safe-State Ladder

We define four states and one hazard (Fig. 1). **S0** is nominal operation. **S1**, balanced standstill, is the legged analogue of monitored standstill: the robot is stationary, balance control is active, and power is required to hold the state. **S2**, seated or braced, is a statically stable posture in which the robot's center of mass lies within a support polygon that does not depend on actuator torque; power is required to reach S2 but not to hold it, so S2 is the first state from which de-energization is safe. **S3**, controlled fall, is the last-resort state in which the balance controller has conceded and a fall-mitigation controller shapes direction and impact. **H** is the uncontrolled fall with a person inside the fall zone. Table I characterizes each state.

The ladder resolves an apparent conflict with ISO 13850. The emergency-stop function must be Category 0 or Category 1 [2]. A Category 0 stop on a biped is a direct transition to H. But Category 1 is defined as a controlled stop with power available to achieve the stop, followed by power removal when the stop is achieved [1], and nothing in that definition requires the stop to be achieved in the current posture. If the stop is defined as the attainment of S2, the seated descent followed by de-energization is a Category 1 stop. We denote it **Category 1D** (descend, then de-energize) to make the posture change explicit. Table II sets out how the stop categories behave across the three machine classes.

B. Response Window and Fall Zone

Two geometric quantities bound the problem. The first is the time available to respond once balance is lost. Modeling the robot as an inverted pendulum of effective length L about its stance edge, a small initial deviation θ_0 grows as $\theta(t) = \theta_0 \cosh(t/\tau)$ in the linear regime [13], so the time to reach a threshold angle θ_c beyond which no capture step exists [14] is

$$t_c = \tau \, \operatorname{arccosh}\!\left(\frac{\theta_c}{\theta_0}\right), \qquad \tau = \sqrt{L/g} \tag{1}$$

Relation (1) bounds the combined latency of fall detection, decision and the first corrective action; it is the legged-robot analogue of the response time T in ISO 13855 [28]. The second quantity is the radius around the robot within which a fall can deliver energy to a person. The stance pivot is fixed only after the robot has completed the step in progress, so the robot can translate by its speed v times the pivot-fixing latency t_s; it then rotates about the pivot through its height H plus any extension ℓ_x of arms or payload beyond the head line; and an intrusion and position-uncertainty allowance C is added by analogy with ISO 13855:

$$R_f = v\, t_s + H + \ell_x + C \tag{2}$$

Relation (2) is deliberately conservative: it takes the fall direction as unknown and the full body length as the reach of the fall. A rated fall-direction controller [16] would justify a sector rather than a circle, but only if its coverage is credited under Section III-C.

C. Hazardous-Event Rate

Three paths reach H (Fig. 1). First, a demand for S2 is issued and the transition fails, with probability q_R per demand. Demands arrive from person approach at rate λ_P and from E-stop actuations, detected internal faults and low-reserve conditions at rate λ_E. Second, balance is lost spontaneously — a slip, an external push, an estimator divergence — at rate $\lambda_{B,m}$ in operating mode m, while a person is inside R_f with probability $p_{z,m}$. Third, the fall controller engages but fails to redirect the fall away from the person, which we capture with a coverage factor c_F. The hazardous-event rate is

$$\lambda_H = q_R \left(\lambda_P + p_{z,E}\, \lambda_E\right) + (1 - c_F) \sum_m \lambda_{B,m}\, p_{z,m} \tag{3}$$

The person-approach term carries no occupancy factor because a person-triggered demand is, by construction, issued with a person nearby; we take the conditional probability that the person is inside R_f as one. This is the structural point of the framework. In a fixed-base cell, a protective stop with a passive end state contributes nothing to the hazard rate however often it is demanded. Here, every demand is a trial of a transition that can fail, conducted in the presence of the person the demand was meant to protect. Under a seat-on-approach policy the protective function's demand rate multiplies its own failure probability.

Relation (3) is a hazardous-event rate, not a probability of dangerous failure per hour of a single safety function as ISO 13849-1 defines it. We nevertheless compare it with the PFH bands for the Performance Level required by the risk assessment, because a safety case for a barrier-free robot has to show that the residual rate of the hazardous event is commensurate with that level, whatever the internal partition. We return to this in Section V.

D. Reach-Failure Decomposition

The probability q_R is not a component property. A seated descent fails if the energy available is insufficient (E), if actuator authority is lost in a joint the descent requires (A), if state estimation diverges during the transient (S), if the descent path is obstructed by an object or the floor is outside its friction assumption (P), or if the descent exceeds its time bound (T). Some of these modes can be diagnosed before a demand arrives — an energy-reserve monitor, actuator health checks, encoder plausibility — and a mode diagnosed before demand converts a reach failure into a planned descent, preferably with no person present. With DC_i the pre-demand diagnostic coverage of mode i and q_i its undiagnosed per-demand probability,

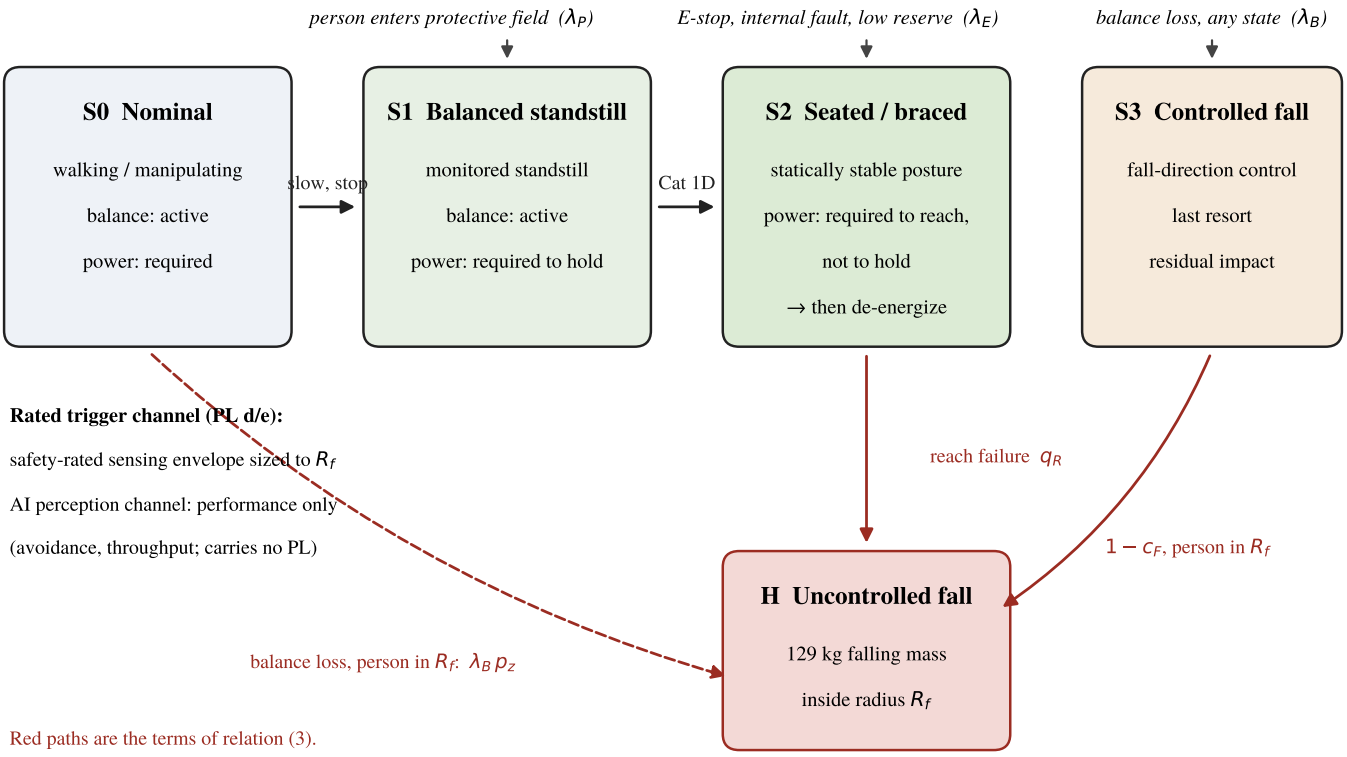


Fig. 1. The safe-state ladder for a dynamically stable robot. Demands arrive from above: person approach (λ_P), E-stop and internal faults (λ_E), and spontaneous balance loss (λ_B). Only S2 is both safe and de-energizable, and it must be reached under power. The red paths into H are the three terms of relation (3): a reach failure during the S1→S2 transition, a balance loss with a person inside the fall radius, and a controlled fall whose direction control does not clear the person.

TABLE I

THE SAFE-STATE LADDER. THE COLUMN THAT MATTERS IS THE LAST: S2 IS THE ONLY STATE THAT IS BOTH SAFE AND SURVIVES DE-ENERGIZATION, AND IT IS THE ONLY ONE WHOSE ENTRY IS A PER-DEMAND TRANSITION THAT CAN FAIL.

State	Posture	Balance control	Power to reach	Power to hold	De-energizable
S0 Nominal	Walking, manipulating	Required	—	Yes	No
S1 Balanced standstill	Standing, stationary	Required	Yes	Yes	No
S2 Seated / braced	Statically stable	During descent only	Yes	No	**Yes**
S3 Controlled fall	Falling, shaped	Fall controller	Yes	—	After impact
H Uncontrolled fall	Falling, unshaped	None	No	—	—

TABLE II

STOP-CATEGORY SEMANTICS BY MACHINE CLASS. UNDER STATIC STABILITY EVERY CATEGORY ENDS UPRIGHT; UNDER ACTIVE STABILITY CATEGORY 0 IS THE HAZARD AND CATEGORY 1 IS SAFE ONLY IF THE DESCENT IS TREATED AS PART OF THE SAFETY FUNCTION.

Category	Fixed-base robot	Statically stable mobile robot	Dynamically stable robot
0 (remove power)	Brakes engage; stops in place	Coasts or brakes; remains upright	**Transition to H**
1 (controlled, then remove)	Decelerate, then remove power	Decelerate, then remove power	Safe only as 1D: reach S2, then remove power
2 (controlled, hold)	Monitored standstill	Monitored standstill	S1: safe while balance holds; not de-energizable
Failure of the stop	Longer stopping distance	Longer stopping distance	**Fall from height**

$$q_R \approx \sum_{i \in \{E,A,S,P,T\}} (1 - DC_i)\, q_i \qquad (4)$$

Relation (4) is the reachability analogue of proof testing. Its practical content is that modes E and A are largely diagnosable, mode S partially, and modes P and T hardly at all, because they depend on the environment at the moment of demand. A safety

concept for a dynamically stable robot should therefore spend its diagnostic effort on E and A and its design effort on making P and T small by construction: a descent posture that needs little floor area and a short, bounded trajectory.

E. Standstill or Seat: The Break-Even Condition

When a person approaches, the robot can hold S1 for the encounter or descend to S2. Holding S1 for an encounter of duration T_e exposes the person to standstill balance loss at rate $\lambda_{B,s}$, of which a fraction c_F is redirected. Descending exposes the person once to q_R, after which S2 is passively stable; we assume the subsequent stand-up is interlocked against an occupied fall zone. Seated descent is the safer proximity response if and only if

$$q_R < q_R^* = \lambda_{B,s}\, T_e\, (1 - c_F) \tag{5}$$

Because standstill balance loss is rare and encounters are short, q_R^* is small, and relation (5) will usually favor standstill. The seated descent then finds its proper role: not a response to proximity, but the procedure by which the robot reaches a de-energizable state when de-energization is actually required — an E-stop, a detected fault, a low reserve, or the end of a shift.

F. Demonstration Burden, Hold-Up Energy and Demand Budget

A claimed q_R must be supported. If n independent descents across the posture and load envelope are executed without failure, the one-sided upper confidence bound at confidence γ satisfies $(1 - q_R)^n \le 1 - \gamma$ [29], so the number of trials required to support a target q_R is approximately

$$n \ge \frac{\ln(1/(1-\gamma))}{q_R} \tag{6}$$

At $\gamma = 0.95$ the numerator is about 3.0, the familiar rule of three. Relation (6) is the legged-robot form of the argument Kalra and Paddock made for automated driving [26]: targets below about 10^{-5} per demand cannot be supported by trial counts alone and require the decomposition of relation (4), with each term argued from its own evidence.

Mode E deserves its own condition because it is the mode most likely to be overlooked. The descent must be completed even when the event that demanded it is a loss of the main power path. Let P_d be the peak electrical power drawn during the descent and T_d its bounded duration. A rated hold-up source, independent of the main battery path and monitored, must satisfy

$$E_{HU} \ge k_s\, P_d\, T_d \tag{7}$$

with a safety factor k_s. The energy is small — the descent lowers the center of mass, and gravity supplies the work — but the electrical draw of holding torque during a controlled eccentric motion is not, and the source must be available at precisely the moment the main source is not.

Finally, rearranging relation (3) gives the demand budget: the largest rate of person-present descents the Performance Level allows once the balance terms have taken their share,

$$\lambda_{P,\max} = \frac{\mathrm{PFH}_{\mathrm{PL}_r} - (1 - c_F) \sum_m \lambda_{B,m}\, p_{z,m}}{q_R} \tag{8}$$

If the numerator is negative, no seat-on-approach policy meets the Performance Level whatever q_R is, and the balance terms alone must be reduced — by speed limits that shrink R_f, by separation monitoring that lowers $p_{z,m}$, or by field evidence that lowers $\lambda_{B,m}$.

IV. Worked Example: A Barrier-Free Industrial Humanoid

A. The Item and the Public Record

We apply the framework to the published parameters of Digit 5 [7]: mass 129 kg, height 1.81 m, reach 2.2 m, repeat-lift payload 22.7 kg, AI-based person monitoring across multiple sensor modalities, an independent safety controller that triggers the response when a person is within an unsafe distance, and three listed responses — avoid, stop, or assume a seated position. The machine is intended for early access in the first half of 2027, including the EU market, where Regulation (EU) 2023/1230 applies from January 20, 2027 and its Annex III essential health and safety requirements apply whether or not a harmonized Type C standard exists [30]. Everything else in this section is our assumption, flagged as such, and none of it is Agility's analysis. The point of the example is the shape of the result, not the decimal places.

B. Applying the Relations

Response window. We assume an effective pendulum length of $L = 1.0$ m, which gives $\tau = 0.319$ s. From an initial deviation of 1° to a capture threshold of 10°, relation (1) gives $t_c = 0.96$ s; the full nonlinear pendulum reaches the floor 1.67 s after a 1° disturbance and 1.16 s after a 5° push. The linear approximation agrees with numerical integration of the nonlinear pendulum to within 0.1 ms at 10°. A fall-detection and response chain therefore has roughly 0.4 to 1.0 s to act, which excludes any architecture in which the fall decision is arbitrated by a remote supervisor.

Fall zone. With an assumed walking speed of 1.5 m/s, a pivot-fixing latency of 0.4 s (one step), the published height of 1.81 m, an extension $\ell_x = 0.39$ m equal to the difference between published reach and height, and $C = 0.2$ m, relation (2) gives $R_f = 3.0$ m. The potential energy released by a full fall about the assumed center-of-mass height is $mgL = 1{,}265$ J.

Demand and balance parameters. For a shared warehouse aisle we assume 12 person approaches per hour into the protective field (one every five minutes) with an encounter duration of 30 s, so the occupancy of the fall zone during standstill is $p_{z,s} = 0.1$. We assume a standstill balance-loss rate $\lambda_{B,s} = 10^{-5}$ /h, a walking balance-loss rate $\lambda_{B,w} = 10^{-3}$ /h, a probability $p_{z,w} = 10^{-3}$ that a person is inside R_f while the robot walks (set by a rated separation-monitoring layer), and a fall-direction coverage $c_F = 0.5$. We assume a development campaign of about 3,000 seated descents without failure, which by relation (6) supports $q_R \le 10^{-3}$ at 95 percent confidence. Table III collects the

TABLE III
WORKED-EXAMPLE PARAMETERS AND RESULTS. PUBLISHED VALUES ARE FROM [7]; ALL OTHERS ARE THE AUTHOR'S ASSUMPTIONS FOR ILLUSTRATION.

Quantity	Value	Source
Mass m / height H / reach	129 kg / 1.81 m / 2.2 m	Published
Pendulum length L; τ	1.0 m; 0.319 s	Assumed; (1)
t_c (1° → 10°); time to floor	0.96 s; 1.67 s	(1); numerical
Speed v; t_s; ℓ_x; C	1.5 m/s; 0.4 s; 0.39 m; 0.2 m	Assumed
Fall radius R_f	3.0 m	(2)
λ_P; T_e; $p_{z,s}$	12 /h; 30 s; 0.1	Assumed
$\lambda_{B,s}$; $\lambda_{B,w}$; $p_{z,w}$; c_F	10^{-5} /h; 10^{-3} /h; 10^{-3}; 0.5	Assumed
Demonstrated q_R (3,000 trials)	$\leq 10^{-3}$	(6)
λ_H, seat on approach	1.2×10^{-2} /h	(3)
λ_H, standstill on approach	1.0×10^{-6} /h	(3)
Break-even q_R^*	4.2×10^{-8}	(5)
Trials to support q_R^*	7.1×10^7	(6)
E-stop q_R target; trials	5×10^{-5}; 6.0×10^4	(3), (6)
Hold-up energy E_{HU}	2.7 kJ (0.75 Wh)	(7)

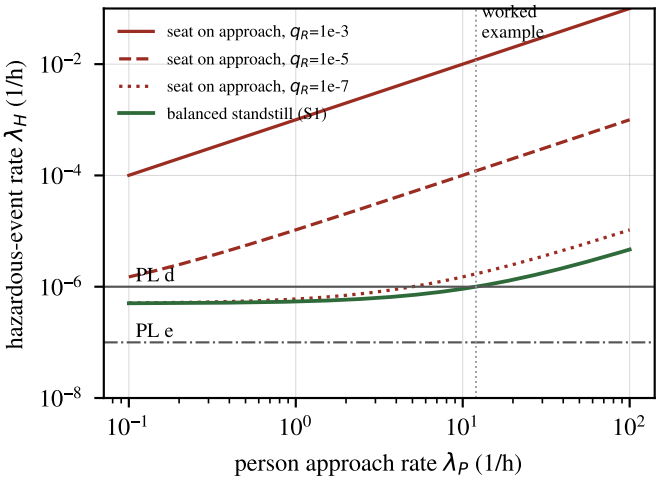


Fig. 2. Hazardous-event rate of relation (3) against person-approach rate for a seat-on-approach policy at three values of reach-failure probability and for a balanced-standstill policy, using the Table III assumptions. The standstill curve rises only when encounters begin to overlap; the seat curves rise linearly with every approach.

parameters and results.

Seat on approach. If every approach triggers a seated descent, the first term of relation (3) is $q_R \lambda_P = 10^{-3} \times 12 = 1.2 \times 10^{-2}$ per hour. The balance terms add $(1 - c_F)\lambda_{B,w} p_{z,w} = 5 \times 10^{-7}$ per hour while walking; the standstill term vanishes because the robot does not stand during encounters. The total, 1.2×10^{-2} per hour, exceeds the PL d ceiling of 10^{-6} per hour [3] by four orders of magnitude and the PL e ceiling by five. Even if the balance terms were zero, relation (8) would limit person-present descents to one per thousand hours at the demonstrated q_R. Fig. 2 plots both policies against the approach rate.

Standstill on approach. If the robot holds S1 for each encounter, the standstill term is $(1 - c_F)\lambda_{B,s} p_{z,s} = 0.5 \times 10^{-5} \times 0.1 = 5 \times 10^{-7}$ per hour, and the walking term adds another 5×10^{-7}. The total, 1.0×10^{-6} per hour, sits on the PL d boundary and an order of magnitude above PL e. The approach-time policy is therefore not a matter of preference: standstill is four orders of magnitude better than seating on this example.

Break-even. Relation (5) gives $q_R^* = 10^{-5} \times (30/3600) \times 0.5 = 4.2 \times 10^{-8}$. Seating would be the safer proximity response only below that value, which relation (6) shows would take about 71 million failure-free descents to support by trial. For a long encounter — a technician working beside the robot for an hour — the threshold rises only to 5×10^{-6}, still well below what can be demonstrated. On these numbers, seating is never the right response to a person.

E-stop. An emergency stop is different: the robot must reach a de-energized state, and S2 is the only one available. Assume E-stop actuations at $\lambda_E = 0.01$ per hour with a person nearby, so $p_{z,E} = 1$, and allocate half of the PL d budget to this term. Relation (3) then requires $q_R \leq 5 \times 10^{-7}/0.01 = 5 \times 10^{-5}$, and relation (6) requires about 60,000 failure-free descents to support it. That is demanding but feasible if a fault-injection campaign covering the E and A modes of relation (4) carries part of the evidence. The demonstrated 10^{-3} is not enough: an E-stop policy that seats at 10^{-3} contributes 10^{-5} per hour by itself.

Hold-up energy. Assuming a peak descent draw of 1.5 kW over a bounded 1.5 s descent with $k_s = 1.2$, relation (7) gives $E_{HU} = 2.7$ kJ, about 0.75 Wh. This is well within the capacity of a supercapacitor bank, which makes the requirement inexpensive to meet. What it adds is the obligation to monitor that store and to treat it as part of the E-stop function.

C. What the Example Says About the Published Architecture

Agility's published architecture already contains the elements the framework calls for: a separation of AI perception from an independent safety controller, and three graded responses [7]. The framework adds an ordering among the responses and a specification for each. Avoidance and slowing are performance functions. Balanced standstill is the protective response to proximity, and its integrity rests on $\lambda_{B,s}$, a quantity that has to be measured in the field because no analysis will produce it. The seated descent is the Category 1D stop for E-stop, fault and low-reserve demands, and its integrity rests on q_R, a quantity that has to be decomposed because no test campaign will demonstrate it whole. The architecture should not leave the choice among the three to the learned perception channel. If the unrated channel decides whether to sit or stand, the Performance Level of the combined function is undefined [3].

D. Derived Requirements

DS-2 reframes what a previous analysis of this machine treated as a deviation from ISO 13850 — a delay of power removal after an E-stop — as conformance with Category 1, provided the descent is treated as part of the stop. The difference is not cosmetic. A deviation must be argued bilaterally with every notified body.

TABLE IV
REQUIREMENTS DERIVED FROM THE FRAMEWORK AND WORKED EXAMPLE, WITH THE RELATION THAT MOTIVATES EACH AND THE CLAUSE UNDER WHICH IT WOULD BE WRITTEN.

ID	Requirement	From	Anchor
DS-1	The safety concept shall define the safe state of each protective function as a posture (S1 or S2), state whether power is required to reach and to hold it, and identify S2 as the only de-energizable safe state.	Table I	ISO 12100 §5.5; ISO/CD 25785-1
DS-2	The E-stop function shall be implemented as a Category 1D stop: a bounded descent to S2 followed by power removal, with the descent treated as part of the safety function and rated at the E-stop's required Performance Level.	Table II, (3)	ISO 13850 §4.1.5; IEC 60204-1 §9.2.2
DS-3	The proximity response to a person entering the protective field shall be balanced standstill (S1) unless relation (5) is shown to favor S2 with q_R supported by evidence; the selection shall be made by the rated channel, not the AI perception channel.	(5)	ISO 13849-1 (SRP/CS)
DS-4	The rated detection envelope shall enclose the fall radius R_f of relation (2) computed at the maximum permitted speed, and speed shall be limited so that R_f remains inside the envelope with a declared margin.	(2)	ISO 13855 §5; ISO 10218-2:2025
DS-5	The reach-failure probability shall be argued by mode as in relation (4); modes E and A shall have pre-demand diagnostics that initiate a planned descent when the fall zone is unoccupied.	(4)	ISO 13849-1 Annex E
DS-6	A hold-up energy source independent of the main battery path shall satisfy relation (7), shall be continuously monitored, and shall inhibit task assignment when its charge falls below the descent demand with margin.	(7)	ISO 13849-1; ISO/CD 25785-1
DS-7	Field balance-loss events shall be recorded by operating mode, and $\lambda_{B,m}$ shall be re-estimated against the value assumed in relation (3) at each periodic review of the risk assessment.	(3), (8)	ISO 12100 §5.6; EU 2023/1230 Annex III

A Category 1 stop whose controlled phase is rated at the required Performance Level is already within the standard's vocabulary; the remaining work is evidence.

V. DISCUSSION

The hazard rate is not a PFH, and the comparison is a proxy. ISO 13849-1 assigns a PFH to a safety function realized by safety-related parts of a control system [3]. Relation (3) combines a per-demand transition failure, a spontaneous mechanical-control event rate and a human-occupancy probability into the rate of a hazardous event. We compare it with the PL bands because the risk assessment that set $\mathrm{PL_r}$ was conducted for the hazardous event, not for any one function, and the residual rate of that event is what a reviewer will ask about. A stricter treatment would allocate the $\mathrm{PL_r}$ budget across functions explicitly. The direction of every result in Section IV survives that allocation, because the dominant seat-on-approach term exceeds the whole budget by four orders of magnitude.

Balance-loss rates are the weakest input. We have no public field data for $\lambda_{B,s}$ or $\lambda_{B,w}$ on industrial humanoids. Operating-hour figures have been published for deployed humanoids, but fall counts have not, and the DARPA Robotics Challenge record, in which falls were frequent [21], comes from prototypes under competition conditions. The values in Table III are placeholders chosen to be plausible, not estimates. This is why DS-7 exists. The framework is useful precisely because it shows which unknown dominates: under a standstill policy the answer is $\lambda_{B,m}$, and under a seat policy it is q_R.

Fall-direction coverage is not yet creditable. The fall-mitigation literature demonstrates that direction can be shaped [16], [17], [19], [20] but reports outcomes in terms of robot damage, impulse or success rate on simulated or laboratory perturbations, not as a coverage factor with a confidence bound. Our $c_F = 0.5$ is an assumption. If c_F cannot be credited, every balance term doubles, and the standstill policy in Section IV moves from the PL d boundary to twice the PL d ceiling. A fall controller implemented as a learned policy has the same Performance Level problem as learned perception.

Independence and the inverted pendulum. Relation (1) treats the robot as a point-mass pendulum and ignores stepping, which is the main recovery strategy [14]. That is conservative for the response window, since stepping extends it, and non-conservative for the fall zone, since a recovery step can move the pivot. We absorbed the second effect into t_s in relation (2). The terms of relation (3) are also treated as independent, whereas a slippery floor raises $\lambda_{B,m}$ and mode P of q_R together. A dependent-failure analysis of environmental coupling is outside the scope of this paper and is needed before the relations are used quantitatively.

Where the standards lens stops. The framework writes reachability into the vocabulary of ISO 13849-1, IEC 60204-1 and ISO 13850 without changing any of them. It cannot settle the question those standards leave open: what Performance Level a learned perception or fall-control function can carry. Our requirements route around that question by assigning every safety-relevant decision to a rated channel and treating the learned channels as performance layers [25], [24]. That is defensible today. It will become restrictive as learned controllers improve, and the place to resolve it is ISO/CD 25785-1 and its planned integration part, both still in draft [4], [8].

Threats to validity. The worked example is constructed from published launch parameters and stated assumptions, not from

an incident, because no public fall-onto-person incident involving an industrial humanoid has been reported. The 12-per-hour approach rate describes a busy shared aisle and would be much lower in a segregated zone, where seat-on-approach becomes less harmful but still offers no benefit over standstill. The results are most decisive in exactly the deployment the product was launched for.

VI. Conclusion and Future Work

For a dynamically stable robot, the safe state is a destination rather than a default, and a destination can be missed. This paper has expressed that fact in the units machinery safety already uses. A four-rung ladder separates the states that need power to hold from the one that needs power only to reach. A hazardous-event rate shows that each demand for the seated state is a trial of a transition that can fail, conducted next to the person who caused the demand. A break-even condition shows that, on plausible numbers, balanced standstill is the better response to proximity by several orders of magnitude, and that the seated descent belongs to de-energization: the E-stop, the fault, the low reserve. Applied to the published parameters of a 129 kg barrier-free humanoid, a seat-on-approach policy exceeds the PL d budget by four orders of magnitude; a standstill policy sits on its boundary; and the E-stop descent needs a reach-failure probability near 5×10^{-5}, supported by a mode-by-mode argument and a monitored hold-up source.

Three directions follow. First, field balance-loss rates by operating mode are the most valuable missing data in humanoid safety, and the industry should agree on a reporting format before fleets grow. Second, fall-direction control needs a coverage metric with a confidence bound so that it can be credited instead of only demonstrated. Third, the Category 1D interpretation proposed here should be tested against the ISO/CD 25785-1 text as it advances, with the aim of making the descent-then-de-energize stop an explicit provision rather than an argument each manufacturer must win separately.

References

[1] Safety of Machinery — Electrical Equipment of Machines — Part 1: General Requirements, IEC 60204-1:2016+A1:2021, IEC, Geneva, Switzerland. [Online]. Available: https://webstore.iec.ch/publication/26037

[2] Safety of Machinery — Emergency Stop Function — Principles for Design, ISO 13850:2015, ISO, Geneva, Switzerland. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:13850:ed-3:v1:en

[3] Safety of Machinery — Safety-Related Parts of Control Systems — Part 1: General Principles for Design, ISO 13849-1:2023, ISO, Geneva, Switzerland. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:13849:-1:ed-4:v1:en

[4] Robotics — Safety Requirements for Dynamically Stable Industrial Mobile Robots (Legged, Wheeled, or Other Forms of Locomotion) — Part 1: Robots, ISO/CD 25785-1, ISO/TC 299, stage 30.60, Jul. 2026. [Online]. Available: https://www.iso.org/committee/5915511.html

[5] S. Haddadin, A. Albu-Schäffer, and G. Hirzinger, "Requirements for safe robots: Measurements, analysis and new insights," *Int. J. Robot. Res.*, vol. 28, no. 11–12, pp. 1507–1527, 2009, doi: 10.1177/0278364909343970.

[6] Robotics — Safety Requirements — Part 1: Industrial Robots, ISO 10218-1:2025, ISO, Geneva, Switzerland. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:10218:-1:ed-3:v1:en

[7] Robotics 24/7 Staff, "Agility Robotics debuts Digit 5 humanoid robot for cooperatively safe work at scale," *Robotics 24/7*, Sep. 15, 2026. [Online]. Available: https://www.robotics247.com/article/agility-robotics-debuts-digit-5-humanoid-robot-for-cooperatively-safe-work-at-scale

[8] IEEE Humanoid Study Group, "Humanoid robots: Report of future standards development," IEEE Standards Association, Piscataway, NJ, USA, Sep. 2025. [Online]. Available: https://www.therobotreport.com/wp-content/uploads/2025/09/IEEE-Humanoid-Report-of-Future-Standards-Development.pdf

[9] N. E. Sanders, E. Şener, and K. B. Chen, "Robot-related injuries in the workplace: An analysis of OSHA Severe Injury Reports," *Appl. Ergon.*, vol. 121, Art. no. 104324, Nov. 2024, doi: 10.1016/j.apergo.2024.104324.

[10] U.S. Occupational Safety and Health Administration, "Robotics: Hazard recognition and standards," Safety and Health Topics. [Online]. Available: https://www.osha.gov/robotics

[11] Industrial Trucks — Safety Requirements and Verification — Part 4: Driverless Industrial Trucks and Their Systems, ISO 3691-4:2023, ISO, Geneva, Switzerland. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:3691:-4:ed-2:v1:en

[12] M. Vukobratović and B. Borovac, "Zero-moment point — Thirty five years of its life," *Int. J. Humanoid Robot.*, vol. 1, no. 1, pp. 157–173, Mar. 2004, doi: 10.1142/S0219843604000083.

[13] S. Kajita, F. Kanehiro, K. Kaneko, K. Yokoi, and H. Hirukawa, "The 3D linear inverted pendulum mode: A simple modeling for a biped walking pattern generation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Maui, HI, USA, 2001, vol. 1, pp. 239–246, doi: 10.1109/IROS.2001.973365.

[14] J. Pratt, J. Carff, S. Drakunov, and A. Goswami, "Capture point: A step toward humanoid push recovery," in *Proc. 6th IEEE-RAS Int. Conf. Humanoid Robots*, Genoa, Italy, 2006, pp. 200–207, doi: 10.1109/ICHR.2006.321385.

[15] S. Kalyanakrishnan and A. Goswami, "Learning to predict humanoid fall," *Int. J. Humanoid Robot.*, vol. 8, no. 2, pp. 245–273, Jun. 2011, doi: 10.1142/S0219843611002484.

[16] A. Goswami, S.-K. Yun, U. Nagarajan, S.-H. Lee, K. Yin, and S. Kalyanakrishnan, "Direction-changing fall control of humanoid robots: Theory and experiments," *Auton. Robots*, vol. 36, no. 3, pp. 199–223, Mar. 2014, doi: 10.1007/s10514-013-9343-2.

[17] V. C. V. Kumar, S. Ha, and C. K. Liu, "Learning a unified control policy for safe falling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Vancouver, BC, Canada, 2017. [Online]. Available: https://arxiv.org/abs/1703.02905

[18] R. Subburaman, D. Kanoulas, N. Tsagarakis, and J. Lee, "A survey on control of humanoid fall over," *Robot. Auton. Syst.*, vol. 166, Art. no. 104443, Aug. 2023, doi: 10.1016/j.robot.2023.104443.

[19] Z. Meng, T. Liu, L. Ma, Y. Wu, R. Song, W. Zhang, and S. Huang, "SafeFall: Learning protective control for humanoid robots," arXiv:2511.18509, Nov. 2025. [Online]. Available: https://arxiv.org/abs/2511.18509

[20] Z. Xu, Y. Li, K.-Y. Lin, and S. X. Yu, "Unified humanoid fall-safety policy from a few demonstrations," arXiv:2511.07407, Nov. 2025. [Online]. Available: https://arxiv.org/abs/2511.07407

[21] C. G. Atkeson *et al.*, "What happened at the DARPA Robotics Challenge finals," in *The DARPA Robotics Challenge Finals: Humanoid Robots to the Rescue*, M. Spenko, S. Buerger, and K. Iagnemma, Eds. Cham, Switzerland: Springer, 2018, pp. 667–684, doi: 10.1007/978-3-319-74666-1_17.

[22] J. A. Marvel and R. Norcross, "Implementing speed and separation monitoring in collaborative robot workcells," *Robot. Comput.-Integr. Manuf.*, vol. 44, pp. 144–155, Apr. 2017, doi: 10.1016/j.rcim.2016.08.001.

[23] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, Nov. 2018, doi: 10.1016/j.mechatronics.2018.02.009.

[24] P. A. Lasota, T. Fong, and J. A. Shah, "A survey of methods for safe human-robot interaction," *Found. Trends Robot.*, vol. 5, no. 4, pp. 261–349, 2017, doi: 10.1561/2300000052.

[25] J. Guiochet, M. Machin, and H. Waeselynck, "Safety-critical advanced robots: A survey," *Robot. Auton. Syst.*, vol. 94, pp. 43–52, Aug. 2017, doi: 10.1016/j.robot.2017.04.004.

[26] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?," *Transp. Res. Part A: Policy Pract.*, vol. 94, pp. 182–193, Dec. 2016, doi: 10.1016/j.tra.2016.09.010.

[27] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE Int. J. Transp. Saf.*, vol. 4, no. 1, pp. 15–24, Apr. 2016, doi: 10.4271/2016-01-0128.

[28] Safety of Machinery — Positioning of Safeguards With Respect to the Approach Speeds of Parts of the Human Body, ISO 13855:2010, ISO, Geneva, Switzerland. [Online]. Available: `https://www.iso.org/obp/ui/#iso:std:iso:13855:ed-2:v1:en`

[29] M. Rausand and A. Høyland, *System Reliability Theory: Models, Statistical Methods, and Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2004, doi: 10.1002/9780470316900.

[30] Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery, *Off. J. Eur. Union*, L 165, pp. 1–102, Jun. 29, 2023. [Online]. Available: `https://eur-lex.europa.eu/eli/reg/2023/1230/oj`