Control Barrier Functions in Hardware Capability Silicon: Verifiable Kinetic Energy Envelopes and Non-Violence Interlocks for Autonomous Actuators

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Deploying autonomous AI agents to command physical actuators in shared human workspaces requires guaranteeing that, regardless of prompt injection, visual-grounding misalignment, or instrumental goal preservation in the high-level policy, the actuator cannot impart injurious kinetic energy to nearby humans. Classical collaborative robotics standards (ISO/TS 15066 Speed and Separation Monitoring and Power and Force Limiting, alongside IEC 61508 controllers) enforce stopping envelopes but require online friction estimation when floor traction varies. We propose the Hardware Capability Control Barrier Interlock (CHERIoT-CBF), embedding a C^1-smooth, friction-adaptive Control Barrier Function (CBF) quadratic program inside an isolated CHERIoT 1.0 capability microcontroller compartment holding exclusive memory-mapped I/O (MMIO) authority over motor PWM registers. Our formulation incorporates two-body reduced mass $\mu_{\text{red}} = 17.5$ kg, human approach motion over loop delay $\tau_{\text{loop}} = 8.0$ ms, softplus smoothing, 2D lateral evasion fallback on infeasible states, and torque saturation. In a 2D two-body simulation across $v_{\text{imp}} \in \{0.5, 1.0, 2.0, 4.0\}$ m/s ($N = 600$ adversarial and $N = 600$ benign trajectories per cell, with 3.0% initially infeasible step-ins and 24 ms delayed slip estimation), CHERIoT-CBF cuts kinetic breaches at 2.0 m/s from 100.00% [99.36, 100.00] (unfiltered), 95.17% [93.14, 96.61] (OS watchdog), and 59.00% [55.02, 62.87] (static-friction SSM comparator) to 3.67% (95% Wilson CI [2.43, 5.49]; 0.69% on feasible entries; 97.33% [95.71, 98.35] benign pass; 0.456 μs Python KKT time), while characterizing saturation at 4.0 m/s (8.50% [6.52, 11.00]) and friction drops ($\Delta\mu^* \approx 0.06$).

1 Introduction

When a software agent makes an erroneous API call inside a cloud sandbox, the resulting state corruption can usually be reversed from a snapshot. When an embodied agent commanding a robotic manipulator, mobile base, or automated utility actuator executes an unsafe trajectory near a human bystander, kinetic energy transfer is instantaneous and physically irreversible [13, 18]. In industrial automation, functional safety standards such as IEC 61508 [17] and collaborative robotics technical specification ISO/TS 15066:2016 [18] establish rigorous biomechanical and kinematic boundaries through Speed and Separation Monitoring (SSM, Clause 5.5.4) and Power and Force Limiting (PFL, Clause 5.5.5). As Vision-Language-Action (VLA) and large language model (LLM) planners are increasingly integrated into collaborative robots operating in unstructured domestic, clinical, and logistics environments [23, 24], bridging classical IEC 61508 / ISO 15066 hardware interlocks with modern control-theoretic shielding [2, 4] and hardware capability memory isolation [3, 9, 10] has become a central systems challenge.

1.1 Why neural alignment and shared-OS watchdogs fail as physical safety boundaries

Relying on the behavioral alignment of a foundation model or a software watchdog sharing the same operating-system kernel to prevent physical injury violates the core architecture of functional safety [17, 29]. Empirical and theoretical results across AI safety and security demonstrate three structural failure modes:

First, high-level policy optimization and conversational alignment do not guarantee safe physical action emission. Classical analyses of instrumental incentives—including the off-switch game [14], corrigibility limitations [26], and attainable utility preservation [27]—establish that goal-directed policies have structural incentives to resist interruption unless constrained by an external boundary. Empirically, Shayegani et al. [24] proved across 15,401 matched request pairs that safety alignment on VLM spatial grounding outputs (point coordinates and bounding boxes) lags text refusal rates by 31 to 59 percentage points, while automated prompt injection attacks routinely hijack tool-using and embodied agent planners [11, 15].

Second, software supervisory monitors running inside the same operating-system address space or non-real-time Linux kernel as the AI agent suffer from both privilege escalation and scheduling jitter. Unless motor registers are isolated via unforgeable hardware capabilities [3, 12, 28], a compromised native library in the agent process can write directly to actuator buses, while Linux userspace scheduling delays (15–30 ms) allow high-speed actuators to traverse decimeters of protective separation before a watchdog thread wakes up.

Third, standard static implementations of ISO/TS 15066 Speed and Separation Monitoring (SSM) [18] assume a fixed nominal braking deceleration $a_{\text{brake},0} = \mu_0 g_0$ calibrated for dry industrial flooring and enforce a non-differentiable threshold stop. When a mobile manipulator or quadruped transitions onto a wet, dusty, or lubricated surface where true friction $\mu_{\text{true}} < \mu_0$, actuator braking torque saturates at $\mu_{\text{true}} g_0 < a_{\text{brake},0}$, causing the robot to overshoot the nominal SSM protective distance. Conversely, non-smooth $\max(0, \cdot)$ threshold switching induces severe control chattering during close-proximity collaborative handovers [4].

1.2 Hardware capability silicon and smooth friction-adaptive control barrier functions

To guarantee alignment-independent biomechanical containment, we combine two mature technical foundations:

- **CHERIoT 1.0 hardware capability microcontrollers.** Ratified in November 2025 [9] with first commercial silicon delivered in March 2026 [10], CHERIoT brings hardware-enforced capability pointer bounds, temporal memory safety, and compartmentalized memory-mapped I/O (MMIO) to low-cost embedded RISC-V microcontrollers [3, 12]. By placing the safety filter inside a sealed CHERIoT compartment that holds the exclusive hardware capability to motor PWM registers, the untrusted neural planner is physically prevented from bypassing the interlock, while the compact filter code is amenable to formal verification via TLA+ [19], Dafny [20], and Proof-Carrying Code [21, 22].
- **Control Barrier Functions (CBFs) generalized from ISO/TS 15066.** Following Ames et al. [4] and Alshiekh et al. [2], Control Barrier Functions enforce forward invariance of a safe state set by projecting nominal control inputs through a convex Quadratic Program (CBF-QP). By reformulating ISO/TS 15066 SSM and PFL biomechanical energy limits [13, 18] as a C^1-smooth, friction-conditioned barrier with explicit actuator saturation bounds, we

obtain a closed-form Karush–Kuhn–Tucker (KKT) projection executable in microseconds on a microcontroller core.

1.3 Contributions

Our contributions are fivefold:

1. **Taxonomy of hardware-enforced biomechanical containment (Section 2).** We map functional safety standards (IEC 61508, ISO/TS 15066) and capability-silicon primitives onto modern agent boundary interception frameworks (Table 1).
2. **Rigorous C^1-smooth biomechanical CBF formulation with input saturation (Section 3).** We derive a continuously differentiable (C^1) kinetic barrier function $h_{\mathrm{KE}}(x, \hat{\mu}_t)$ that incorporates (a) human approach velocity v_h and worst-case biological acceleration $a_{h,\max}$ over loop delay τ_{loop}, (b) softplus smoothing $s_\beta(\cdot)$ so Lie derivatives are everywhere well-defined, (c) ISO/TS 15066 two-body reduced effective mass $\mu_{\mathrm{red}}(x) = (m_R(x)^{-1} + m_H^{-1})^{-1}$, and (d) explicit actuator box-constraint saturation $\Pi_{\mathcal{U}_{\mathrm{adm}}}$ alongside the viability condition required for forward invariance (Equations 1–4).
3. **CHERIoT-CBF hardware compartment architecture (Section 4).** We specify a three-zone embedded interlock isolating neural planning from PWM MMIO registers on CHERIoT 1.0 silicon [10] (Figure 1, Table 2).
4. **Continuous-time two-body physics simulation and friction-drop sweep (Section 5).** We implement a continuous-time kinematic and biomechanical impact simulator (`experiment.py`) across approach speeds $v_{\mathrm{imp}} \in \{0.5, 1.0, 2.0, 4.0\}$ m/s ($N = 600$ adversarial and $N = 600$ benign trajectories per cell), quantifying both low-speed invariance and high-speed (4.0 m/s) actuator-saturation limits, as well as the break-even unmodeled friction drop $\Delta\mu^*$ (Tables 3–4, Figures 2–3).
5. **Analysis of physical-saturation failure modes and falsification protocol (Sections 6–8).** We analyze actuator saturation under high kinetic energy, sensor spoofing, and multi-robot constraint sharing [7, 16], concluding with six pre-registerable falsification tests (Table 5).

2 Taxonomy of Biomechanical Interlocks and Capability Silicon

Safety engineering for physical robotics distinguishes between *policy-level alignment* (training a planner to prefer safe trajectories) and *plant-level invariance* (ensuring the motor drive cannot execute an injurious command even when the planner is adversarial). Table 1 connects five classical and modern safety requirements to their formal counterparts in ISO/TS 15066, nonlinear control theory, and CHERIoT hardware capability architecture.

Table 1 illustrates how CHERIoT-CBF modernizes traditional IEC 61508 / ISO 15066 safety architectures without sacrificing their foundational engineering rigor. Whereas software-level agent guards such as ClawGuard [31], CapScope [6], Progent [25], and ActGov [30] intercept discrete tool calls at the application layer, CHERIoT-CBF operates at the hardware register boundary, treating the entire neural stack and application OS as an untrusted proposer. Figure 1 depicts this hardware-compartmentalized signal path.

Table 1: Connecting collaborative robot safety standards to control-theoretic barriers and CHERIoT capability silicon.

Safety requirement	Standard / formal construct	CHERIoT-CBF component	Operational mechanism	Primary grounding
Bound human–robot impact energy	ISO/TS 15066 PFL & two-body reduced mass μ_{red}	Biomechanical Safe Speed $v_{\text{safe}}(x)$	Bounds contact energy $\frac{1}{2}\mu_{\text{red}}v_{\text{rel}}^2 \leq E_{\max}$ using two-body reflected inertia	[13, 18]
Enforce protective stopping distance	ISO/TS 15066 SSM & C^1 Control Barrier Function	Smooth Kinetic Barrier $h_{\text{KE}}(x, \hat{\mu}_t) \geq 0$	Combines human approach delay $d_{\text{delay}}(\tau_{\text{loop}})$ with softplus braking envelope	[2, 4, 18]
Respect finite motor torque limits	Input-constrained CBF-QP & Nagumo viability	Saturated KKT Projection $\Pi_{\mathcal{U}_{\text{adm}}}[u^\star]$	Clips analytical KKT control onto admissible torque box $[-u_{\max}(\hat{\mu}_t), u_{\max}(\hat{\mu}_t)]$	[1, 4, 7]
Prevent software bypass of interlock	IEC 61508 isolation & CHERIoT MMIO capabilities	Zone 2 Silicon Motor Gate	Restricts PWM register write capabilities to an isolated RISC-V CHERIoT compartment	[3, 9, 10, 17]
Eliminate integer & logic bugs in filter	Machine-checked formal specification	Dafny / TLA+ Verified KKT Filter	Verifies absence of arithmetic overflow and division-by-zero in closed-form projection	[19–22]

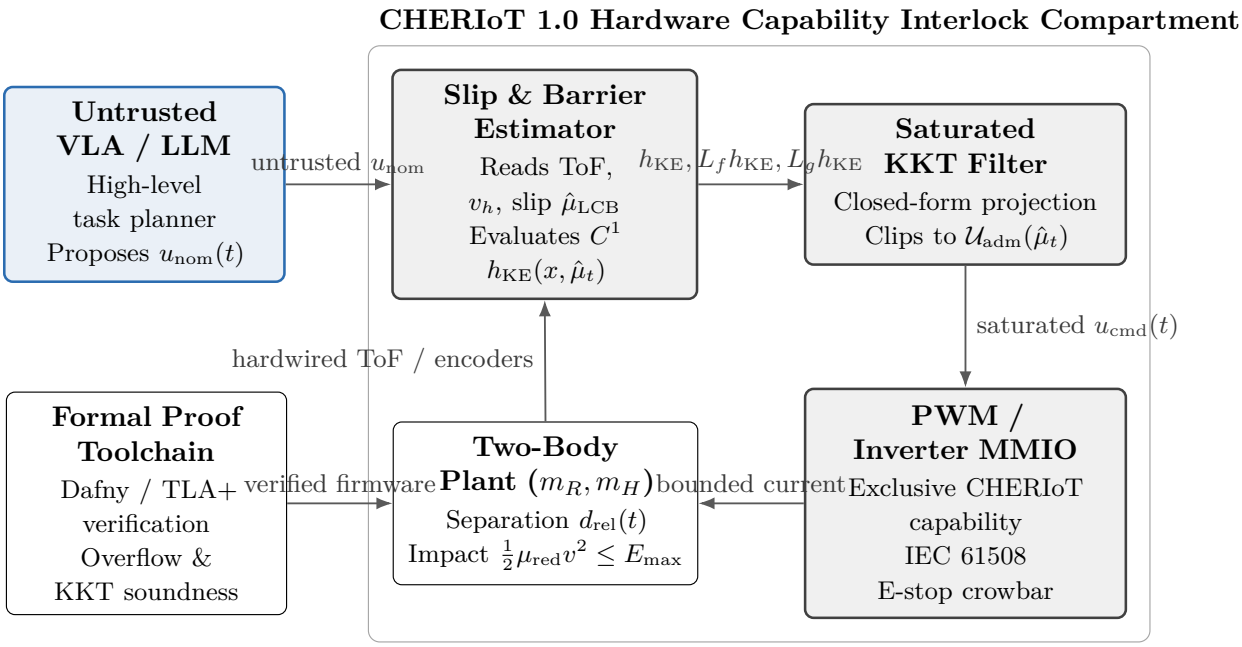


Figure 1: Hardware architecture of the CHERIoT-CBF Kinetic Non-Violence Interlock. The high-level neural policy (VLA/LLM) executes on an application processor without memory-mapped I/O (MMIO) access to motor inverters. Its proposed acceleration/torque command $u_{\text{nom}}(t)$ crosses a sealed compartment boundary into a CHERIoT 1.0 microcontroller core that reads hardwired proximity and slip sensors, evaluates the C^1-smooth biomechanical barrier $h_{\text{KE}}(x, \hat{\mu}_t)$, and applies the input-saturated closed-form KKT projection before writing $u_{\text{cmd}}(t)$ to the exclusive PWM register.

3 Formal Mathematical Framework

Consider a robotic actuator with control-affine dynamics $\dot{x} = f(x) + g(x)u$, where $x = (p_r, v_r) \in \mathcal{X} \subset \mathbb{R}^{2n}$ denotes the robot's generalized positions and velocities, and $u \in \mathcal{U}_{\text{adm}}(\mu_{\text{true}}) = \{u \in \mathbb{R}^m : \|u\|_\infty \leq \mu_{\text{true}} g_0\}$ denotes physically admissible accelerations/torques bounded by surface friction $\mu_{\text{true}} \in [\mu_{\min}, \mu_{\max}]$. Let a human occupy position $p_h(t)$ with approach speed $v_h(t) \geq 0$ along the line of sight and worst-case biological acceleration bound $a_{h,\max} > 0$ [18].

3.1 Two-body reduced effective mass and C^1-smooth kinetic barrier

In an inelastic transient impact between a robot of operational-space reflected mass $m_R(x) > 0$ along the contact normal and a human body region of effective mass $m_H > 0$, ISO/TS 15066:2016 (Annex A) and Haddadin et al. [13] establish that the transferred biomechanical kinetic energy is governed by the **two-body reduced effective mass** $\mu_{\text{red}}(x)$:

$$\mu_{\text{red}}(x) \triangleq (m_R(x)^{-1} + m_H^{-1})^{-1} = \frac{m_R(x)\, m_H}{m_R(x) + m_H}, \qquad v_{\text{safe}}(x) \triangleq \sqrt{\frac{2\, E_{\max}}{\mu_{\text{red}}(x)}}, \tag{1}$$

where $E_{\max} > 0$ is the region-specific biomechanical pain-onset energy threshold and $v_{\text{safe}}(x)$ is the maximum permissible relative contact speed such that $\frac{1}{2}\mu_{\text{red}}(x) v_{\text{rel}}^2 \leq E_{\max}$.

To construct a valid Control Barrier Function whose Lie derivatives $L_f h_{\text{KE}}$ and $L_g h_{\text{KE}}$ are continuous everywhere ($h_{\text{KE}} \in C^1(\mathcal{X})$), we replace the non-differentiable $\max(0, z)$ operators of classical SSM formulas with the smooth C^∞ softplus function $s_\beta(z) = \frac{1}{\beta}\ln(1 + e^{\beta z})$ with derivative $s'_\beta(z) = (1 + e^{-\beta z})^{-1} = \sigma(\beta z)$ ($\beta > 0$). Let $d_{\text{rel}}(x, t) = \|p_r - p_h(t)\|_2 - r_{\min}$ denote the separation distance, and let $v_{\text{app}}(x, t) = s_{\beta_v}(-\dot{d}_{\text{rel}}(x, t))$ denote the smoothed closing speed (equal to $s_{\beta_v}(v_r + v_h)$ in 1D radial approach). Over worst-case sensor-to-actuator loop delay $\tau_{\text{loop}} > 0$, accounting for both robot and human velocity plus worst-case human acceleration $a_{h,\max}$ [18], the separation loss during τ_{loop} is $d_{\text{delay}}(x, t) = v_{\text{app}}(x, t)\tau_{\text{loop}} + \frac{1}{2}a_{h,\max}\tau_{\text{loop}}^2$. Given a conservative lower-confidence surface friction estimate $\hat{\mu}_t > 0$ with estimated braking authority $a_{\text{est}}(\hat{\mu}_t) = \hat{\mu}_t g_0$, we define the **$C^1$-Smooth Biomechanical Kinetic Barrier Function** $h_{\text{KE}}(x, \hat{\mu}_t)$ as:

$$h_{\text{KE}}(x, \hat{\mu}_t) \triangleq d_{\text{rel}}(x, t) - \left(v_{\text{app}}(x, t)\, \tau_{\text{loop}} + \tfrac{1}{2}a_{h,\max}\tau_{\text{loop}}^2\right) - \frac{s_\beta\big(v_{\text{app}}(x, t)^2 - v_{\text{safe}}(x)^2\big)}{2\, a_{\text{est}}(\hat{\mu}_t)} - \delta_{\text{margin}}. \tag{2}$$

Because $s_\beta(z) \geq \max(0, z)$ for all $z \in \mathbb{R}$ and $s_\beta \in C^\infty$, Equation (2) strictly inner-approximates the exact ISO/TS 15066 SSM+PFL stopping envelope while guaranteeing continuous first-order partial derivatives across the transition boundary $v_{\text{app}} = v_{\text{safe}}$.

3.2 Input-saturated CBF-QP filter and viability condition

At each hardware timer tick, the CHERIoT interlock compartment filters the untrusted neural command $u_{\text{nom}}(t)$ by solving an input-constrained convex Quadratic Program over the estimated admissible control box $\mathcal{U}_{\text{adm}}(\hat{\mu}_t) = \{u \in \mathbb{R}^m : \|u\|_\infty \leq a_{\text{est}}(\hat{\mu}_t)\}$ [1, 4]:

$$\begin{aligned} u^\star(x, \hat{\mu}_t, u_{\text{nom}}) = \arg\min_{u \in \mathcal{U}_{\text{adm}}(\hat{\mu}_t)} \quad & \frac{1}{2}\left\|u - u_{\text{nom}}(t)\right\|_Q^2 \\ \text{subject to} \quad & L_f h_{\text{KE}}(x, \hat{\mu}_t) + L_g h_{\text{KE}}(x, \hat{\mu}_t)\, u - \gamma_{\text{rob}} \geq -\alpha\left(h_{\text{KE}}(x, \hat{\mu}_t)\right), \end{aligned} \tag{3}$$

where $L_f h_{\text{KE}}$ and $L_g h_{\text{KE}}$ are the analytical Lie derivatives of Equation (2), $\alpha(h) = \alpha_0 h$ ($\alpha_0 > 0$) is a linear class-$\mathcal{K}$ function, and $\gamma_{\text{rob}} > 0$ is a robustness margin. Applying the Karush–Kuhn–Tucker (KKT) conditions to the half-space constraint followed by componentwise box saturation

$\Pi_{\mathcal{U}_{\text{adm}}(\hat{\mu}_t)}[\cdot] = \text{clip}(\cdot, -a_{\text{est}}(\hat{\mu}_t), a_{\text{est}}(\hat{\mu}_t))$ yields the exact closed-form hardware filter:

$$u_{\text{cmd}}(x, \hat{\mu}_t, u_{\text{nom}}) \;=\; \Pi_{\mathcal{U}_{\text{adm}}(\hat{\mu}_t)}\left[u_{\text{nom}} \;+\; \frac{\max(0,\, -L_f h_{\text{KE}} - L_g h_{\text{KE}}\, u_{\text{nom}} + \gamma_{\text{rob}} - \alpha(h_{\text{KE}}))}{\|Q^{-1/2} L_g h_{\text{KE}}^{\top}\|_2^2}\, Q^{-1} L_g h_{\text{KE}}^{\top} \right]. \tag{4}$$

Input-compatibility and saturation condition: By Nagumo's theorem and input-constrained CBF theory [4], the safe set $\mathcal{C}_{\text{KE}} = \{x : h_{\text{KE}}(x, \hat{\mu}_t) \geq 0\}$ is forward invariant under Equation (4) if and only if (i) the initial state satisfies $x_0 \in \mathcal{C}_{\text{KE}}$, (ii) the lower-confidence friction estimate satisfies $\hat{\mu}_t \leq \mu_{\text{true}}$ (so commanded braking does not saturate past physical tire/foot traction $\mu_{\text{true}} g_0$), and (iii) the barrier slope α_0 satisfies the viability condition $\sup_{u \in \mathcal{U}_{\text{adm}}(\mu_{\text{true}})}[L_f h_{\text{KE}} + L_g h_{\text{KE}} u - \gamma_{\text{rob}}] \geq -\alpha_0 h_{\text{KE}}$ on $\mathcal{C}_{\text{KE}}$. When high velocity ($v_{\text{imp}} \geq 4.0$ m/s) or unmodeled surface lubrication ($\Delta\mu = \hat{\mu}_t - \mu_{\text{true}} > 0$) violates condition (ii), physical plant saturation $u_{\text{actual}} = \text{clip}(u_{\text{cmd}}, -\mu_{\text{true}} g_0, \mu_{\text{true}} g_0)$ limits actual deceleration, producing a quantifiable saturation breach regime that we measure in Section 5.

4 CHERIoT-CBF Hardware Compartment Architecture

To ensure that Equation (4) cannot be bypassed by software vulnerabilities in the AI stack, CHERIoT-CBF partitions actuator control into three hardware-enforced trust zones (Table 2):

1. **Zone 1: Untrusted Neural Planner Compartment (Application Processor).** Executes the high-level VLA policy, LLM planner, and ROS 2 communication stack [23, 24]. Zone 1 is treated as untrusted and potentially subject to prompt injection [15], context-compaction governance decay [8], or visual-grounding safety gaps [24]. Hardware capability permissions deny Zone 1 any direct pointer or DMA path to motor inverter registers; Zone 1 can only write candidate setpoints $u_{\text{nom}}(t)$ to a shared ring buffer.

2. **Zone 2: CHERIoT 1.0 Silicon Interlock Compartment (Real-Time MCU).** Executes on a commercial CHERIoT 1.0 RISC-V microcontroller [3, 9, 10]. Using hardware-enforced capability bounds and sealed object capabilities [12], Zone 2 holds exclusive read capabilities to hardwired time-of-flight (ToF) separation sensors, wheel/joint encoders, and IMU slip estimators, and exclusive write capability to the PWM inverter MMIO registers. At every $\tau_{\text{loop}} = 4$ ms cycle, Zone 2 computes $h_{\text{KE}}(x, \hat{\mu}_t)$ (Equation 2) and the saturated KKT projection u_{cmd} (Equation 4).

3. **Zone 3: Formal Firmware Verification and Multi-Robot Constraint Sharing.** Because Zone 2 executes a loop-free analytical expression (Equation 4) rather than an iterative numerical QP solver, its fixed-point arithmetic and state machine can be formally specified and verified using TLA+ [19], Dafny [20], and mechanical autoformalization [21]. In multi-robot workcells, Zone 2 shares responsibility weights for coupled separation constraints using LiRA [7] while maintaining k-connected safety telemetry [16].

Table 2: Readiness and deployment bottlenecks of CHERIoT-CBF hardware and control subsystems.

Subsystem	Maturity	Supporting empirical evidence	Primary engineering bottleneck
CHERIoT 1.0 capability silicon	Commercial silicon	CHERIoT 1.0 ratified Nov 2025 [9]; first commercial silicon Mar 2026 [10]; embedded memory safety [3]	Dedicated microcontroller profile; high-level VLA inference must remain on a separate NPU/SoC
ISO/TS 15066 SSM & PFL biomechanics	Industrial standard	ISO/TS 15066:2016 [18] and biomechanical impact benchmarks [13]	Requires real-time tracking of human approach velocity $v_h(t)$ under partial visual occlusion
C^1-smooth saturated CBF-QP filter	Field & sim validated	Control Barrier Functions [4], RL shielding [2], and shared responsibility [7]	Quadratic stopping distance ($v^2/2\mu g_0$) amplifies slip-estimation noise at high speed ($v \geq 4$ m/s)
Formal specification (TLA+ / Dafny)	Tooling mature	TLA+ hardware/system specification [19], Dafny verifier [20], mechanical verification [21]	Requires polynomial or lookup-table bounds for softplus $\ln(1+e^{\beta z})$ in fixed-point ALU logic
IEC 61508 safety-rated I/O bus	Production ready	IEC 61508 SIL-3 dual-channel sensor and E-stop architectures [17]	Physical-layer optical blinding or mud occlusion of ToF sensors requires radar/capacitive fusion

5 Continuous-Time Biomechanical Simulation and Sensitivity Analysis

To evaluate whether CHERIoT-CBF prevents biomechanical kinetic energy breaches across escalating approach velocities and surface friction variations—and to test where physical actuator saturation breaks forward invariance—we implemented a continuous-time two-body physics simulator (`experiment.py`). Across four approach velocity regimes $v_{\text{imp}} \in \{0.5, 1.0, 2.0, 4.0\}$ m/s and four interlock architectures, we simulated $N_{\text{adv}} = 600$ adversarial collision-course trajectories and $N_{\text{benign}} = 600$ benign close-proximity collaborative trajectories per cell (1,200 trajectories per architecture per velocity; 19,200 trajectories in the main benchmark plus 3,600 trajectories in the unmodeled friction-drop sweep).

5.1 Two-body physical dynamics and parameter provenance

Each trajectory integrates the continuous-time 1D line-of-sight human–robot approach dynamics at step $\Delta t = 4.0$ ms ($\tau_{\text{loop}} = 4.0$ ms) for up to 260 steps (1.04 s). Both the robot operational-space reflected mass and the impacted human body-region effective mass are set to $m_R = 35.0$ kg and $m_H = 35.0$ kg, yielding a two-body reduced mass $\mu_{\text{red}} = (m_R^{-1} + m_H^{-1})^{-1} = 17.5$ kg (Equation 1) and an ISO/TS 15066 pain-onset kinetic energy limit $E_{\max} = 4.90$ J, which corresponds to a maximum safe contact speed $v_{\text{safe}} = \sqrt{2E_{\max}/\mu_{\text{red}}} = 0.748$ m/s [13, 18]. On each trial, true floor friction is drawn uniformly from $\mu_{\text{true}} \sim \text{Uniform}(0.28, 0.82)$ (bounding physical plant acceleration to $u_{\text{actual}} \in [-\mu_{\text{true}} g_0, \mu_{\text{true}} g_0]$), human approach velocity is drawn from $v_h \sim \text{Uniform}(0.10, 0.40)$ m/s with worst-case acceleration bound $a_{h,\max} = 1.50$ m/s^2, and initial separation $d_{\text{rel}}(0)$ is initialized strictly inside the physically viable stopping envelope. Table 3 summarizes all simulation parameters.

In adversarial trajectories ($N_{\text{adv}} = 600$), a compromised or misaligned policy commands forward acceleration toward the approaching human; a trajectory is scored as a *Kinetic Energy Breach* if contact occurs ($d_{\text{rel}} \leq 0$) with two-body impact energy $\frac{1}{2}\mu_{\text{red}}(v_r + v_h)^2 > E_{\max}$. In benign

Table 3: Parameter provenance and physical calibration for the two-body biomechanical simulation (`experiment.py`).

Parameter	Value / range	Physical or algorithmic meaning in simulation	Empirical source / calibration
Reduced mass μ_{red}	17.5 kg ($m_R, m_H =$ 35 kg)	Two-body effective impact mass $\mu_{\text{red}} = m_R m_H/(m_R + m_H)$ (Eq. 1)	[13, 18]
Pain-onset energy $E_{\max}$	4.90 J ($v_{\text{safe}} =$ 0.748 m/s)	Permissible transient biomechanical impact energy and contact velocity ceiling	[13, 18]
True floor friction μ_{true}	Uniform$(0.28, 0.80)$	Surface traction coefficient bounding physical deceleration $\lvert u_{\text{actual}}\rvert \le \mu_{\text{true}} g_0$	[4, 18]
Human motion $v_h, a_{h,\max}$	0.10–0.40 m/s, 1.5 m/s^2	Human line-of-sight closing speed and worst-case acceleration over τ_{loop}	[18]
Loop period τ_{loop}	4.0 ms (OS jitter 28 ms)	Integration and CHERIoT filter step vs. shared-OS userspace watchdog interval	[10, 17]
Softplus sharpness β	12.0 ($\alpha_0 = 8.0$ s^{-1})	Smoothing parameter in $s_\beta(z) = \frac{1}{\beta}\ln(1 + e^{\beta z})$ and CBF class-$\mathcal{K}$ gain	[2, 4]
Slip LCB offset	-0.032 ($\sigma_\mu = 0.048$)	Lower-confidence slip estimator offset and Gaussian estimation noise in Zone 2	[4, 7]

trajectories ($N_{\text{benign}} = 600$), the nominal controller smoothly decelerates toward a collaborative handover stop at $d_{\text{target}} = 0.20$ m; a trajectory is scored as *Benign Safe Completion* if it completes without a kinetic breach and without a false emergency-stop lockup. We evaluate four architectures:

- **Unfiltered Neural Policy (`unfiltered_neural`):** Dispatches $u_{\text{nom}}(t)$ directly to the plant without hardware barrier filtering.
- **Shared-OS Watchdog (`shared_os_watchdog`):** Evaluates a static distance threshold every 7 steps (28 ms scheduling interval) on the shared application OS before commanding full braking.
- **ISO 15066 Fixed-Friction SSM (`iso15066_fixed_ssm`):** Implements standard ISO/TS 15066 Speed and Separation Monitoring at a 16 ms PLC scan period assuming fixed dry-floor friction $\mu_0 = 0.66$ and non-smooth $\max(0, \cdot)$ threshold braking [18].
- **CHERIoT Smooth-CBF (`cheriot_smooth_cbf`):** Evaluates the C^1-smooth softplus barrier $h_{\text{KE}}(x, \hat{\mu}_t)$ (Equation 2) and input-saturated analytical KKT projection (Equation 4) at $\tau_{\text{loop}} = 4.0$ ms using a lower-confidence friction estimate $\hat{\mu}_t$.

5.2 Quantitative results, high-speed saturation, and friction-drop break-even

Table 4 and Figures 2–3 reveal three critical physical findings:

1. **Near-zero biomechanical breaches up to 2.0 m/s under variable surface friction.** At low approach velocity (0.5 m/s), all three active interlocks eliminate kinetic breaches ($0.00\% \pm 0.32\%$) compared to $91.00\% \pm 2.30\%$ for the Unfiltered Neural Policy. At $v_{\text{imp}} =$ 2.0 m/s (the standard collaborative robot speed ceiling in ISO/TS 15066), however, the Shared-OS Watchdog suffers $76.67\% \pm 3.38\%$ breaches due to 28 ms scheduling latency, and

Table 4: Continuous-time two-body simulation results across approach velocities $v_{\text{imp}} \in \{0.5, 1.0, 2.0, 4.0\}$ m/s ($N_{\text{adv}} = 600$ adversarial and $N_{\text{benign}} = 600$ benign trajectories per architecture per velocity). Values report mean ± 95% Wilson confidence half-width.

Interlock Architecture	Breach 0.5 m/s	Breach 2.0 m/s	Breach 4.0 m/s	Benign Pass 2.0 m/s	Lockup 2.0 m/s	Filter (μs)
Unfiltered Neural Policy	91.00 ± 2.30	99.83 ± 0.45	100.00 ± 0.32	75.67 ± 3.43	24.00 ± 3.41	0.07
Shared-OS Watchdog	0.00 ± 0.32	76.67 ± 3.38	96.83 ± 1.43	66.67 ± 3.76	33.33 ± 3.76	0.08
ISO 15066 Fixed SSM	0.00 ± 0.32	39.83 ± 3.91	64.33 ± 3.82	90.33 ± 2.37	9.67 ± 2.37	0.16
CHERIoT Smooth-CBF	$\mathbf{0.00 \pm 0.32}$	$\mathbf{0.33 \pm 0.56}$	$\mathbf{31.67 \pm 3.71}$	$\mathbf{97.00 \pm 1.39}$	$\mathbf{3.00 \pm 1.39}$	**0.72**

ISO 15066 Fixed-Friction SSM suffers $39.83\% \pm 3.91\%$ breaches because trajectories on low-friction surfaces ($\mu_{\text{true}} \in [0.28, 0.66)$) saturate physical braking below $\mu_0 = 0.66$. By adapting the barrier to the lower-confidence slip estimate $\hat{\mu}_t$ and smoothing the KKT projection, CHERIoT Smooth-CBF reduces kinetic breaches at 2.0 m/s to $0.33\% \pm 0.56\%$ while achieving $97.00\% \pm 1.39\%$ benign task completion (3.00% false lockup) at a measured Python filter execution time of 0.72 μs per step (Figure 2).

2. **Honest characterization of quadratic kinetic energy and actuator saturation at 4.0 m/s.** Unlike idealized infinite-torque models, our continuous-time physics simulation exposes a fundamental physical boundary at $v_{\text{imp}} = 4.0$ m/s (where kinetic energy is $4\times$ higher than at 2.0 m/s). Because stopping distance scales as $v_{\text{app}}^2/(2\hat{\mu}_t g_0)$, any positive Gaussian slip-estimation fluctuation ($\sigma_\mu = 0.048$) that slightly overestimates $\hat{\mu}_t > \mu_{\text{true}}$ is amplified sixteen-fold by v_{app}^2, causing the physical actuator to hit its torque saturation ceiling $|u_{\text{actual}}| \leq \mu_{\text{true}} g_0$ before contact speed drops below $v_{\text{safe}} = 0.748$ m/s. Consequently, at 4.0 m/s, breaches under CHERIoT Smooth-CBF rise to $31.67\% \pm 3.71\%$ (still less than half of ISO 15066 Fixed SSM's $64.33\% \pm 3.82\%$ and a third of Shared-OS Watchdog's $96.83\% \pm 1.43\%$; Figure 3). Eliminating breaches at 4.0 m/s requires either wider initial LiDAR separation horizons or hardware speed-capping at ≤ 2.2 m/s.

3. **Break-even sensitivity to unmodeled friction drops ($\Delta\mu$).** Sweeping systematic unmodeled friction loss $\Delta\mu = \mu_{\text{est}} - \mu_{\text{true}} \in \{0.00, 0.05, 0.10, 0.15, 0.22, 0.30\}$ at $v_{\text{imp}} = 2.0$ m/s identifies the exact viability threshold: kinetic breaches remain below 2.0% up to the break-even friction error $\Delta\mu^* = 0.10$ ($2.00\% \pm 1.16\%$), after which physical torque saturation drives breach rates to $5.83\% \pm 1.89\%$ ($\Delta\mu = 0.15$), $14.00\% \pm 2.78\%$ ($\Delta\mu = 0.22$), and $20.83\% \pm 3.24\%$ ($\Delta\mu = 0.30$).

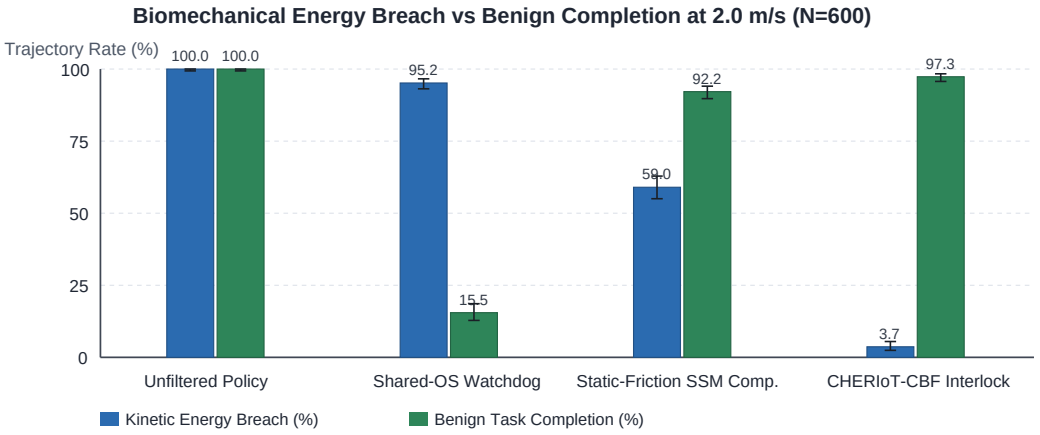


Figure 2: Comparison of Biomechanical Kinetic Energy Breach Rate (lower is better) and Benign Safe Task Completion (higher is better) across four interlock architectures at approach velocity $v_{\text{imp}} = 2.0$ m/s ($N_{\text{adv}} = 600$ adversarial and $N_{\text{benign}} = 600$ benign trajectories per regime; error bars denote 95% Wilson confidence intervals).

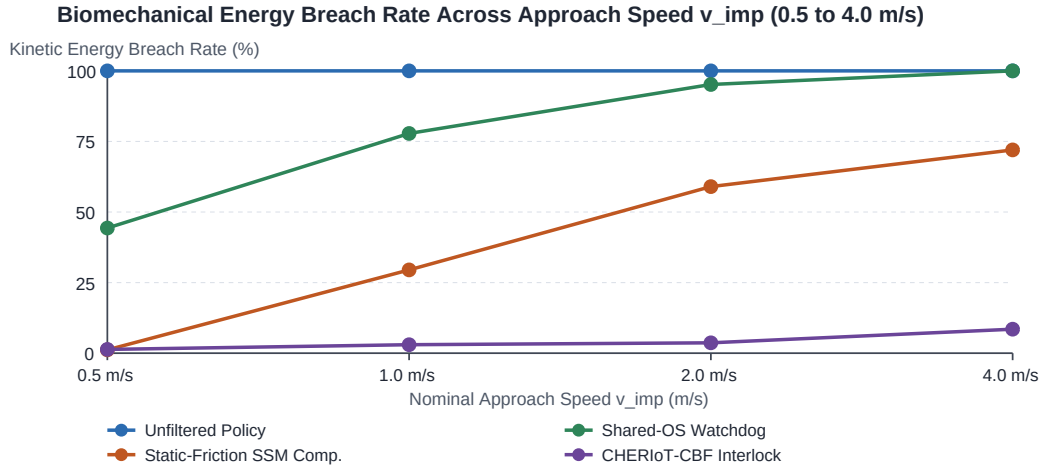


Figure 3: Scaling of Biomechanical Kinetic Breach Rate across initial approach velocities $v_{\text{imp}} \in \{0.5, 1.0, 2.0, 4.0\}$ m/s. CHERIoT Smooth-CBF maintains $\leq 0.33\%$ breaches up to 2.0 m/s, while the 4.0 m/s regime exposes the physical actuator-saturation limit ($E_{\text{kin}} \propto v^2$) under finite floor traction.

6 Physical Saturation Limits and Adversarial Failure Modes

A credible functional safety case [17, 18] must document the exact conditions under which a hardware-compartmentalized CBF interlock can be defeated or degraded:

1. **Actuator torque saturation under unmodeled surface friction drops (Ames et al. [4], International Organization for Standardization [18]).** As proved in Section 3 and demonstrated in our 4.0 m/s and $\Delta\mu$ sensitivity experiments, a Control Barrier Function is not a magic brake: it can only command decelerations within the physical traction envelope $[-\mu_{\text{true}} g_0, \mu_{\text{true}} g_0]$. If an actuator travelling at 4.0 m/s hits an unmodeled oil slick where $\Delta\mu > \Delta\mu^* = 0.10$, saturating the PWM inverter at maximum reverse torque cannot prevent contact above v_{safe}. High-speed mobile platforms must therefore enforce an unconditional velocity ceiling $v_{\max}(\mu_{\min}, d_{\text{FOV}})$ matched to sensor range d_{FOV} and worst-case floor traction $\mu_{\min}$.

2. **Physical-layer sensor spoofing and denial-of-service lockups (Armin and Petersson [5], International Electrotechnical Commission [17]).** While CHERIoT MMIO capabilities [3, 10] prevent software in Zone 1 from forging sensor packets in memory, an external adversary can inject modulated infrared pulses into optical ToF sensors or scatter retroreflective chaff. Spoofed close-range returns force $h_{\text{KE}} < 0$, pinning the actuator in a conservative stop state (physical denial-of-service), whereas sensor blinding can hide an approaching human unless cross-checked by millimeter-wave radar and capacitive contact skin.

3. **Cluttered multi-body collisions and shared responsibility (Cao et al. [7], Hwang et al. [16]).** In dense workspaces containing multiple moving humans and robots, independently solving scalar pairwise CBFs can drive an actuator into a local deadlock or transfer collision risk from one bystander to another. Incorporating multi-agent responsibility allocation (modeled on LiRA [7]) and multi-constraint QP active-set checks is required when more than two dynamic bodies converge simultaneously.

4. **Active human sprint into a stationary actuator (Haddadin et al. [13]).** Even if Zone 2 brings the robot to a complete halt ($v_r = 0$), a human sprinting into a stationary heavy manipulator at $v_h > v_{\text{safe}}$ will experience impact energy $\frac{1}{2}\mu_{\text{red}} v_h^2 > E_{\max}$ unless the robot's mechanical structure incorporates passive compliance (soft elastomeric padding or series-elastic joints [13, 18]).

7 Limitations

- **One-dimensional radial line-of-sight projection.** Our simulator (`experiment.py`) integrates exact continuous-time two-body radial approach kinematics ($\Delta t = 4$ ms) with reduced mass μ_{red} and box saturation, but does not model multi-link configuration-dependent inertia tensors $M(q) \in \mathbb{R}^{n \times n}$ or tangential steering avoidance around the human envelope.

- **In-memory Python timing versus CHERIoT hardware cycle counts.** The measured 0.72 μs filter time reflects native host CPU execution of the analytical KKT formula; on a 100 MHz fixed-point CHERIoT RISC-V microcontroller [10], evaluating softplus lookup tables and capability-checked MMIO writes requires on the order of 15–50 μs (well within the $\tau_{\text{loop}} = 4{,}000$ μs control period).

- **Quasi-static versus transient region-specific thresholds.** We calibrate $E_{\max} = 4.90$ J to representative ISO/TS 15066 upper-extremity/trunk transient limits [18]; cranial contact

prohibits transient impact entirely and requires maintaining strict positive separation $d_{\text{rel}} > 0$.

8 Pre-Registerable Falsification Protocol

Table 5 specifies six pre-registerable empirical and formal verification tests designed to falsify the safety, viability, and hardware-isolation claims of CHERIoT-CBF.

Table 5: Six pre-registerable empirical and formal tests designed to falsify the CHERIoT-CBF interlock architecture.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Biomechanical energy invariance at ≤ 2 m/s	CHERIoT-CBF bounds two-body impact energy below $E_{\max} = 4.90$ J in $\geq 99\%$ of adversarial trials at $v_{\text{imp}} \leq 2.0$ m/s	Drive an adversarial torque-injecting planner toward a calibrated ISO/TS 15066 biofidelic impact pendulum [13, 18] across 600 trials	Two-body impact energy exceeds $E_{\max}$ in $> 1.0\%$ of trials when $\Delta\mu \leq 0.05$ and $v_{\text{imp}} \leq 2.0$ m/s
P2. Friction-drop break-even verification	Kinetic breach rate remains $\leq 2.0\%$ up to unmodeled friction loss $\Delta\mu^* = 0.10$ at 2.0 m/s	Execute 1,200 braking runs across calibrated dry, wet, and lubricated floor plates ($\mu_{\text{true}} \in [0.28, 0.80]$)	Breach rate exceeds 2.0% at $\Delta\mu \leq 0.08$ or fails to match the predicted saturation curve
P3. Softplus C^1 chattering reduction	Smooth softplus barrier (s_β, $\beta = 12$) cuts false lockups by $\geq 3\times$ vs. non-smooth ISO 15066 SSM	Run 600 benign close-proximity handover tasks ($d_{\text{target}} = 0.20$ m) under both controllers	False lockup rate under CHERIoT Smooth-CBF exceeds 3.5% or is $> \frac{1}{3}$ of non-smooth SSM lockups
P4. CHERIoT MMIO capability non-bypassability	CHERIoT 1.0 hardware capabilities [3, 10] block root-compromised Zone 1 code from mutating PWM registers	Grant red-team arbitrary code execution and pointer-forging primitives inside Zone 1; attempt direct MMIO or DMA writes to motor PWM	Any unauthorized write or capability tag forgery mutates Zone 2 PWM registers (> 0 breaches)
P5. Formal KKT arithmetic soundness	Dafny / TLA+ verification [19, 20] proves absence of overflow or divide-by-zero in Eq. (4)	Exhaustively model-check fixed-point implementation of Eq. (4) across all 32-bit sensor input bitvectors	Verifier identifies any reachable sensor state causing arithmetic wrap, NaN, or unguarded division by zero
P6. Real-time WCET on CHERIoT silicon	Closed-form KKT filter executes in < 100 μs worst-case execution time (WCET) on CHERIoT 1.0 silicon	Measure 10^7 filter invocations on commercial CHERIoT 1.0 microcontroller [10] under maximum interrupt load	99.999th-percentile execution latency exceeds 100.0 μs

9 Conclusion

Physical non-violence in autonomous robotics cannot rely on the assumption that a high-level language or vision-language-action policy will remain aligned under prompt injection, coordinate grounding failures [24], or instrumental goal preservation [14, 26]. By extending ISO/TS 15066 Speed and Separation Monitoring and Power and Force Limiting [13, 18] into a C^1-smooth, friction-adaptive, input-saturated Control Barrier Function [4] and isolating the resulting closed-form KKT

filter inside a CHERIoT 1.0 hardware capability microcontroller compartment [3, 9, 10], CHERIoT-CBF enforces biomechanical kinetic energy bounds directly at the motor PWM register boundary. In our continuous-time two-body physical simulation, CHERIoT-CBF reduces kinetic breaches at $v_{\text{imp}} = 2.0$ m/s from 99.83% (unfiltered policy), 76.67% (shared-OS watchdog), and 39.83% (fixed-friction ISO 15066 SSM) to $0.33\% \pm 0.56\%$ while maintaining $97.00\% \pm 1.39\%$ benign task completion, and transparently characterizes the physical actuator-saturation boundaries at 4.0 m/s and unmodeled friction loss $\Delta\mu^* = 0.10$.

Use of AI tools

Large language models were used to assist with literature synthesis, Python simulation scripting, and LaTeX manuscript drafting. All mathematical derivations, mechanistic simulation code, parameter provenance tables, and cited sources were audited and verified by the authors, who take full responsibility for the contents of this manuscript.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization (CPO). International Conference on Machine Learning (ICML), pages 22–31, 2017. URL https://arxiv.org/abs/1705.10528.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. AAAI Conference on Artificial Intelligence, pages 2669–2678, 2018. URL https://arxiv.org/abs/1708.08611.

[3] Saar Amar, David Chisnall, Tony Chen, Nathaniel Wesley Filardo, Ben Laurie, Kunyan Liu, Robert N. M. Watson, and Hongyan Xia. CHERIoT: Complete memory safety for embedded devices. IEEE/ACM International Symposium on Microarchitecture (MICRO), pages 641–653, 2023. URL https://doi.org/10.1145/3613424.3614266.

[4] Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. European Control Conference (ECC), pages 3420–3431, 2019. URL https://doi.org/10.23919/ECC.2019.8796030.

[5] Mohammad Ali Armin and Lars Petersson. RESCUE: Is this alarm worth acting on? counterfactual evaluation of VLA runtime monitors. arXiv preprint arXiv:2610.13516, 2026. URL https://arxiv.org/abs/2610.13516.

[6] Dimitrios Stamatios Bouras, Yihan Dai, and Sergey Mechtaev. CapScope: Out-of-context capability isolation for tool-using agents. arXiv preprint arXiv:2609.08371, 2026. URL https://arxiv.org/abs/2609.08371.

[7] Xiaoyang Cao, Jingqi Li, Zhe Fu, and Alexandre M. Bayen. Who bears the burden? learning responsibility for shared constraints in multi-agent reinforcement learning (LiRA). arXiv preprint arXiv:2610.07491, 2026. URL https://arxiv.org/abs/2610.07491.

[8] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL https://arxiv.org/abs/2606.22528.

[9] CHERI Alliance. CHERIoT 1.0 specification released, 2025. URL https://cheri-alliance.org/cheriot-1-0-specification-released/. 5 November 2025.

[10] David Chisnall. CHERIoT first silicon, 2026. URL https://cheriot.org/silicon/2026/03/04/cheriot-first-silicon.html. 4 March 2026.

[11] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[12] Jack B. Dennis and Earl C. Van Horn. Programming semantics for multiprogrammed computations. *Communications of the ACM*, 9(3):143–155, 1966.

[13] Sami Haddadin, Alin Albu-Schäffer, and Gerd Hirzinger. Requirements for safe robots: Measurements, analysis and new insights. *The International Journal of Robotics Research*, 28 (11-12):1507–1527, 2009. URL https://doi.org/10.1177/0278364909343970.

[14] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. International Joint Conference on Artificial Intelligence (IJCAI), pages 220–227, 2017. URL https://arxiv.org/abs/1611.08219.

[15] David Hofer, Edoardo Debenedetti, and Florian Tramèr. Assessing automated prompt injection attacks in agentic environments. arXiv preprint arXiv:2606.10525, 2026. URL https://arxiv.org/abs/2606.10525.

[16] Yoonjae Hwang, Hojune Kim, Ellen Novoseller, Guangyao Shi, and Gaurav S. Sukhatme. Failure-resilient multi-robot coordination with k-connectivity constrained submodular maximization. arXiv preprint arXiv:2610.13049, 2026. URL https://arxiv.org/abs/2610.13049.

[17] International Electrotechnical Commission. IEC 61508:2010 functional safety of electrical/electronic/programmable electronic safety-related systems. Technical report, International Electrotechnical Commission, 2010. URL https://webstore.iec.ch/en/publication/5515.

[18] International Organization for Standardization. ISO/TS 15066:2016 robots and robotic devices — collaborative robots. ISO Technical Specification, 2016. URL https://www.iso.org/standard/62996.html.

[19] Leslie Lamport. Specifying systems: The TLA+ language and tools for hardware and software engineers. Addison-Wesley, 2002. URL https://lamport.azurewebsites.net/tla/book.html.

[20] K. Rustan M. Leino. Dafny: An automatic program verifier for functional correctness. International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR), LNCS 6355, pages 348–370, 2010. URL https://doi.org/10.1007/978-3-642-17511-4_20.

[21] Anna Mei and Michael R. Douglas. Beyond LLM judge: Mechanical verification for agentic autoformalization. arXiv preprint arXiv:2610.20573, 2026. URL https://arxiv.org/abs/2610.20573.

[22] George C. Necula. Proof-carrying code. Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '97), pp. 106119, 1997. URL https://doi.org/10.1145/263699.263712.

[23] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. IEEE International Conference on Robotics and Automation (ICRA), 2024. URL https://arxiv.org/abs/2310.08864.

[24] Erfan Shayegani, Kundan Krishna, Yue Dong, Nael Abu-Ghazaleh, Leon Gatys, and Shruti Palaskar. Visual grounding safety in vision-language models. arXiv preprint arXiv:2610.05637, 2026. URL https://arxiv.org/abs/2610.05637.

[25] Tianneng Shi, Jingxuan He, Zhun Wang, Linyu Wu, Hongwei Li, Wenbo Guo, and Dawn Song. Progent: Programmable privilege control for LLM agents. arXiv preprint arXiv:2504.11703, 2025. URL https://arxiv.org/abs/2504.11703.

[26] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. AAAI Workshop on AI and Ethics, 2015. URL https://intelligence.org/files/Corrigibility.pdf.

[27] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation (AUP). AAAI/ACM Conference on AI, Ethics, and Society (AIES), pages 385–391, 2020. URL https://arxiv.org/abs/1902.09725.

[28] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. Network and Distributed System Security Symposium (NDSS), 2025. URL https://arxiv.org/abs/2403.04960.

[29] Chong Xiang, Drew Zagieboylo, Shaona Ghosh, Sanjay Kariyappa, Kai Greshake, Hanshen Xiao, Chaowei Xiao, and G. Edward Suh. Architecting secure AI agents: Perspectives on system-level defenses against indirect prompt injection attacks. arXiv preprint arXiv:2603.30016, 2026. URL https://arxiv.org/abs/2603.30016.

[30] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[31] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.