Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 09 | Sep-2026
ISSN (O) 3107-6696

A Context-Aware Approach to Personal Security in Smartwatches: An Interaction Design Framework for High-Stress Safety Systems

Pratishtha Srivastava, Miss Ruchi Gaur

[World University of Design] · [pratishthasrivastava25@wud.ac.in]

Abstract - Commercial smartwatch emergency features are almost uniformly *reactive*: they require the wearer to locate, press, and hold a physical button, or to navigate a touch target on a display of 34–45 mm diagonal, in precisely the circumstances where the underlying motor and attentional resources are least available. Laboratory evidence on acute stress shows that threat-type stressors degrade fine force control by a substantial margin while leaving gross force production intact, that usable visual field area contracts under cognitive load, and that threat raises blink frequency such that the wearer is intermittently not looking at anything at all. A reactive interface therefore fails at the moment of need, while the automatic-detection alternative fails in the opposite direction: published evaluations of commercial wrist-worn fall detection report sensitivity of 77% with a 1.7% per-event false-positive rate in induced-fall protocols, and as low as 4.7% sensitivity in populations outside the tuning distribution — and deep models trained on laboratory falls collapse to 0.37 precision on real-world data.

This paper proposes an interaction design framework that treats threat recognition not as a classification verdict but as a *graded state* negotiated between the system and the wearer. We contribute: (1) a **sensor fusion matrix** specifying physiological, environmental, and behavioural evidence vectors together with their documented failure modes and the fusion level at which each is combined; (2) a **4-State Transition Architecture** (Ambient Baseline, Micro-Verification, Silent Beacon, Critical Broadcast) in which each upward transition is gated by either accumulated evidence or an explicit low-cognition user act, and each state specifies a bounded interaction budget; (3) a **low-dexterity modality ladder** grounded in published performance envelopes for vibrotactile pattern identification, encumbered target acquisition, wrist-rotation gestures, casing squeeze, and bezel input; and (4) a **covert state management** model in which the visually deniable presentation is the default rather than an option, motivated by evidence that the adversary is frequently a device-authenticated, co-present intimate. We specify an evaluation protocol with stress induction, adversarial covertness role-play, and metrics spanning Time-to-Acknowledge, false-positive rate, NASA-TLX, and SUS, and close with five design heuristics for safety-critical wearable interfaces.

Keywords: wearable computing; personal safety; context-aware computing; interaction design; acute stress; vibrotactile interaction; covert interfaces; mixed-initiative systems

1 Introduction and Background

1.1 The reactive premise and why it fails

Every widely deployed smartwatch personal-safety feature shares a design premise: that at the moment of danger the wearer will recognise the danger, form the intention to summon help, select the correct control, and execute a deliberate motor act — typically a sustained press of a side button, sometimes combined with a confirmatory drag on the touchscreen. The premise is attractive because it places the human firmly in control and because it makes the false-positive rate a function of user intent rather than of classifier performance. It is also, we argue, empirically unsound as a *sole* pathway, because each of its four steps is selectively impaired by the physiological state that the interface is designed to serve.

Consider the motor step first. Noteboom, Fleshner, and Enoka [49] had 29 participants perform a precision pinch-grip task under either mental arithmetic or unpredictable electric shock. Both stressors elevated arousal markers — heart rate, systolic blood pressure, electrodermal activity — by between 5% and 80%. Only the shock stressor degraded grip steadiness, and it did so by approximately 130%. The companion study [48] established the dose-response: steadiness impairment scaled with both stressor intensity and trait anxiety, while *mean* motor output was unchanged. This dissociation is the single most consequential finding for safety-interface design, and it cuts two ways. It tells us that arousal alone is not a sufficient predictor of motor degradation — an interface cannot infer "the user cannot press accurately" from heart rate alone. And it tells us that what degrades is *fine force regulation*, not force generation. A wearer under threat can still swing an arm, clench a fist, or rotate a wrist through a large angle; what they cannot reliably do is modulate a 2 N pinch or land a fingertip within a 6 mm target.

Second, the selection step. Nieuwenhuys and Oudejans [44] had police officers shoot an identical exercise against a non-threatening opponent and against an opponent returning fire with marking cartridges. Under high anxiety, accuracy fell, officers acted faster, they crouched to reduce their target profile — and they blinked significantly more often, increasing cumulative eyes-closed time. For a wrist display this is not a marginal observation. It means the assumption behind every glanceable safety UI — that the wearer can look at the watch when prompted — holds only intermittently under threat. The wearer's gaze is additionally committed to

the threat itself: Nieuwenhuys, Cañal-Bruland, and Oudejans [46] found that threat shifted officers' *action thresholds* (they fired earlier) without measurably changing their perceptual distance judgments, consistent with an attentional rather than perceptual mechanism. Nieuwenhuys and Oudejans' integrated model [45] locates these effects at three operational levels — attentional, interpretational, and behavioural — and it is the attentional level that governs our design: the interface must assume that attention is captured by the threat, not available to the device.

Third, the perceptual step. Easterbrook's cue-utilisation hypothesis [24] predicts progressive restriction of the range of cues used as emotional arousal rises, and Staal's review for NASA [63] traces the subsequent consolidation of that account, while cautioning that different stressors produce dissociable performance signatures rather than one general decrement. The quantitative anchor is Rantanen and Goldberg [54], who mapped visual fields by kinetic perimetry along 24 meridians under three levels of concurrent cognitive load: field area shrank 7.74% from light to moderate load and a further 5.81% from moderate to heavy, leaving 86.41% of the light-load field under heavy load ($F(2,39) = 15.21$, $p < 0.001$). Critically, the contraction was *anisotropic*, most pronounced in the upper field within roughly 20° either side of the vertical apex. Dirkin [22] supplies the necessary corrective: a peripherally located task that is *primary* can still be attended to under stress without decrement, indicating that the effect is attentional prioritisation rather than an optical shrinking of the field. The design implication is not "make everything bigger" but "make the safety-critical cue the primary task in a modality the threat is not already occupying."

Fourth, the intention step. A wearer under coercive threat may have an intention that is the opposite of what the interface assumes: not to summon a visible, audible response, but to summon help *without the aggressor noticing*. Karusala and Kumar [31], in the closest prior interaction study of panic buttons, found that women in New Delhi already performed discreet safety signalling — pretending to talk or text in order to appear socially connected — and were openly sceptical that police would respond at all. Their recommendation, notably, was for deliberate activation *friction* (multiple rapid presses, or a long press) to avoid embarrassing false alarms and wasted dispatch. Our framework takes that recommendation seriously while relocating the friction: friction on *escalation* is defensible, friction on *initiation* is not.

Interestingly, the evidence that touch performance degrades under stress is weaker than the folklore suggests, and we state that honestly. Sarsenbayeva et al. [59] induced stress with the Trier Social Stress Test in 24 participants and measured target acquisition on a 5.1-inch phone: touch offset grew from 46.47 px (SD 25.57) at baseline to approximately 50 px under stress ($p < 0.01$), while completion time became marginally *faster* (538.68 ms to 534.77 ms, $p < 0.01$) — a speed–accuracy shift toward faster, sloppier taps rather than a collapse. The stress induction was validated by HF-HRV and STAI. The appropriate reading is that *socially evaluative* stress produces a small offset penalty; it is the threat-type stressors of [49] and [44], and the concurrent physical encumbrance documented by Ng, Brewster, and Williamson [43], that produce large effects. In that study, encumbered walking drove two-handed index-finger accuracy to 48.1% and one-handed thumb accuracy to roughly 31–34% for distant targets, with preferred-thumb targeting error reaching 4.2 mm — a 40% increase over the unencumbered baseline. The authors recommend a minimum 4.2 mm radius for small components used while encumbered and walking. On a wrist display, where Parhi, Karlson, and Bederson's classic thumb study [50] already recommends 9.2 mm targets for discrete tasks, and where Xia, Grossman, and Fitzmaurice [71] showed that removing fingertip occlusion with a finger-mounted stylus cuts smartwatch-scale selection errors by 80%, the arithmetic does not leave room for a multi-target confirmation dialogue.

1.2 Failure modes of the deployed alternative

If reactive interfaces fail under stress, the apparent remedy is automatic detection — the path taken by commercial fall and crash detection. The published evidence indicates that this path substitutes one failure for two.

Sensitivity does not generalise. Brew, Faux, and Blanchard [9] evaluated a smartwatch fall-detection app across 226 induced falls: sensitivity 77%, specificity 99%, false-positive rate 1.7%, false-negative rate 16.4%, overall accuracy 89%, with detection better when the fall occurred on the same side as the watch. Near-falls were detected at 89% sensitivity with zero false positives. These are respectable numbers for a controlled protocol. Abou et al. [1] then tested the Apple Watch Series 5 fall-detection feature across 300 controlled wheelchair falls with 25 participants and measured sensitivity of 4.7% — a 95.3% miss rate — with height, impact magnitude, lower-limb function, and fall direction implicated. A threshold model tuned on ambulatory falls does not transfer to a body that falls differently.

Laboratory precision does not survive deployment. Mauldin et al. [39] trained a deep fall-detection model on wrist accelerometry and reported precision of 0.77–0.79 on laboratory datasets. On the Farseeing corpus of 23 real-world falls, precision for every method collapsed — 0.37 for the deep model, 0.44 for SVM, 0.005 for naive Bayes — while *accuracy* remained 0.99, the canonical class-imbalance trap. In live real-time testing the deep model reached precision 0.64 and recall 0.86.

Specificity is the binding constraint, and it is worse than it looks. Tchuente, Baddour, and Lemaire [64] classified aggressive movements (punching, shoving, slapping, shaking) against confusable benign ones (clapping, waving, handshaking, opening doors, typing) from smartwatch IMU data. With watches on both wrists, kNN with ReliefF reached 99.6% accuracy and 99.8% specificity. With a single dominant-wrist watch — the deployed configuration — accuracy held at 98.1% but specificity fell to 93.6%. This is

the closest published analogue to detecting struggle-like motion from one wrist, and the specificity figure is the one that determines whether the product is usable.

A fourth constraint compounds these: the physiological channel is weakest precisely at the wrist. In the WESAD benchmark [61], binary stress-versus-non-stress classification reached 93.12% from a chest-worn array but only 88.33% from the wrist, and three-class accuracy fell from 80.34% to 76.17%; fusing wrist with chest yielded no improvement over chest alone. The roughly five-point wrist penalty is the ceiling any wrist-only affect inference inherits before motion artefact is even considered.

The reason specificity dominates is a base-rate argument that safety-system designers routinely underweight. Suppose a system evaluates one-minute decision windows across 16 waking hours: 960 windows per day. For nuisance escalations to occur less than once per month, the per-window false-positive probability must be below $1/(960 \times 30) \approx 3.5 \times 10^{-5}$. A per-window false-positive rate of even 0.1% — some sixty times better than the 6.4% complement of the single-wrist specificity reported above — still produces approximately one nuisance escalation per day. The best published single-wrist specificity therefore sits more than three orders of magnitude away from what autonomous escalation to an external responder would require.

The consequences of getting this wrong are documented. In clinical monitoring, Drew et al. [23] logged 2,558,760 unique alarms over 31 days in a single ICU — 187 audible alarms per bed per day — with 88.8% of 12,671 annotated arrhythmia alarms false-positive; the AHRQ review by Woo and Bacon [70] places the false-alarm share across studies between 72% and 99%. Cvach's review [18] frames the resulting desensitisation. Sampson's problem-oriented policing analysis of false burglar alarms [58] documents the same structure on the responder side: automated alarm systems generating response demand that overwhelms the responders they recruit. We note one important dissent: Wickens et al. [69] analysed 495 air-traffic conflict alerts, 45% of them false, and found *no* cry-wolf effect — no non-responses to true alerts and no response delays at higher-false-alarm centres — because controllers retained situational awareness and acted pre-emptively. The lesson is not that false alarms are harmless but that their cost depends on whether the human remains in the loop with enough context to triage. That is precisely the property our State 1 is designed to preserve.

The visibility failure. A third failure mode is orthogonal to accuracy. Commercial emergency features announce themselves: countdown tones, escalating haptics, a siren, a full-screen red interface. Where the threat is a stranger at distance, visibility may deter. Where the threat is co-present and attentive, visibility invites confiscation of the device or escalation of violence. Freed et al. [26], in a study of 89 participants cataloguing 27 attack types, introduced the *UI-bound adversary*: an abuser who needs no technical skill because they are an authenticated, co-present user of the victim's devices and accounts. Chatterjee et al. [12] measured the ecosystem of dual-use applications repurposed for covert partner surveillance, many marketed as safety tools. Bellini [7] found that abusive partners readily admit to GPS location tracking while reframing it as protective family monitoring. Matthews et al. [38] showed that survivors' privacy and security needs invert across the phases of physical control, escape, and life apart. Taken together, this literature establishes that a wearable safety feature is itself a plausible surveillance and coercion vector, and that a design which cannot be concealed from a co-present intimate is, for a substantial population, worse than no design at all. Chen et al. [13] extend the point to alerting aesthetics: trauma-informed computing warns that alarm-styled alerts and interfaces that change unexpectedly can themselves trigger hypervigilance.

1.3 Research questions

We formulate three research questions, each addressed by a distinct contribution of the framework.

RQ1 (Inference and escalation design). How can heterogeneous wrist-sensor evidence — physiological, environmental, and behavioural — be fused into a representation of threat that is expressed as *graded confidence with an associated interaction obligation*, rather than as a binary alarm verdict, and what evidence conditions and dwell times should govern transitions between those grades?

RQ2 (Low-dexterity input under degraded motor and attentional resources). Which interaction modalities remain reliably operable given threat-induced fine-force degradation, anisotropic visual-field contraction, and intermittent gaze availability; and what are their quantified design envelopes — minimum target geometry, time budget, recognition tolerance, and error asymmetry — for the specific acts of *confirming* and *vetoing* an escalation?

RQ3 (Autonomy, deniability, and data minimisation). How can a system that continuously senses the body and its environment preserve user autonomy, plausible deniability, and data minimisation when the adversary may be a co-present, device-authenticated intimate — that is, when the safety feature is simultaneously an attack surface?

2 Related Work

2.1 Wearable safety systems and emergency interaction patterns

The HCI literature on personal safety technology has matured from artefact proposals toward critical empirical evaluation, and the critical turn is where the design constraints are found. Karusala and Kumar [31] remains the reference study for panic-button interaction, contributing both the activation-friction recommendation and the observation that

discreet signalling is already part of users' repertoire. Chordia et al. [15] analysed the Citizen application through interviews and interface analysis and found that it heightened anxiety about safety while steering users toward paid features, extending dark-pattern harm taxonomies to emotional burden and social injustice. The same group [16] then derived five design strategies for reducing *dysfunctional* fear of crime — fear that damages wellbeing without improving safety — which supplies a principled basis for a safety wearable to *suppress* ambient threat signalling rather than maximise it.

On the responder side, Ming et al. [41] analysed LiveSafe deployments across more than 200 universities and found a systematic mismatch between what safety departments expected to receive and what community members actually reported, with anonymity and tip type predicting whether a department responded at all. Park, Kwon, and Lee [51] ran a four-week deployment of communitysourced patrolling with 20 participants and characterised what actually sustained volunteer commitment. Tseng et al. [66] report on eight months of an IPV technology clinic (62 clients, 7 consultants, 18 professionals) and observe that survivors frequently cannot find a safe synchronous moment to seek help — an argument for asynchronous, resumable, low-friction channels over a single well-designed synchronous interaction. Collectively, this work establishes that the weak link in safety systems is rarely the reporting interface; it is the reliability of the response and the social cost of invoking it. A framework that escalates aggressively without modelling that cost optimises the wrong variable.

2.2 Context-aware computing and multi-sensor fusion

Our framework inherits its vocabulary from the context-aware computing tradition. Schilit, Adams, and Want [60] introduced the term together with the taxonomy of proximate selection, automatic reconfiguration, and contextual commands; Dey [20] supplied the operational definition of context as any information characterising the situation of an entity; Abowd et al. [2] distinguished context *capture* from context *interpretation* and enumerated the primary who/what/where/when dimensions; Salber, Dey, and Abowd [57] provided the canonical sensor-abstraction architecture. For fusion mechanics we follow Gravina et al. [28], whose survey of body sensor networks distinguishes data-, feature-, and decision-level fusion and names the energy, synchronisation, and reliability constraints that bound on-device inference; Lara and Labrador [35] and Bao and Intille [5] supply the activity-recognition baselines and the sensitivity of inference to sensor placement.

The critical strand matters more to us than the architectural one. Bellotti and Edwards [8] argued that context-aware systems must be *intelligible* and must keep humans *accountable* for consequential actions — that a system may not act on inferences its users cannot inspect. Lim, Dey, and Avrahami [37] demonstrated empirically that why and why-not explanations improve intelligibility of context-aware systems. Barkhuus and Dey [6] compared personalised, passive context-aware, and actively proactive services and found that users accept reduced control when perceived usefulness is high but remain acutely sensitive to its loss. Horvitz [30] formalised the coupling of automation and direct manipulation under uncertainty, including the expected-utility threshold governing whether a system should act or ask — which is exactly the structure of a dispatch decision. Amershi et al. [3] contribute the validated guideline set, of which *scope services when in doubt*, *support efficient dismissal and correction*, and *convey consequence* map almost one-to-one onto the requirements of automated safety escalation. Kocielnik, Amershi, and Bennett [33] showed experimentally that framing a classifier as high-precision versus high-recall changes user acceptance at *equal* accuracy — directly relevant to how a safety watch should communicate its own error profile.

2.3 Micro-interactions, haptics, and non-visual modalities

Ashbrook [4] defined the microinteraction as an interaction initiated and completed in under four seconds, and the in-the-wild data confirm that smartwatch use is microinteractive by nature: Visuri et al. [68], across 307 users, 798,423 sessions, and 2,801,082 notifications, report a median session length of 5 s, with 65.2% of sessions being "peeks" of 5 s or less, 80% user-initiated, and only 9.4% of notifications producing any interaction session at all. Pizza et al. [53] corroborate with wearable-camera video: across 34 days of recording and 1,009 analysed watch uses, activity is dominated by very brief glances embedded in ongoing tasks. A safety interaction that requires more than a few seconds of sustained attention is not merely stressful under threat; it is unlike anything the wearer does with the device normally.

For non-visual signalling, Brewster and Brown [10] defined *tactons* as structured tactile messages composed from frequency, amplitude, duration, rhythm, body location, and spatiotemporal pattern. The capacity limits are well characterised and sobering: Brown, Brewster, and Purchase [11] found two-parameter tactons (rhythm × roughness) identified at roughly 70%, falling to 48% when body location was added as a third parameter, and recovering to 81% when the number of values on one parameter was reduced. Ternes and MacLean [65] showed that rhythm is the parameter that scales, constructing 84 distinguishable haptic icons, with discrimination driven primarily by note length and unevenness. Lee and Starner [36] studied multi-actuator wrist displays and alert perception under multitasking. Cholewiak and Collins [14] established that localisation on the arm is governed chiefly by proximity to body landmarks and inter-site spacing rather than by frequency. Youn, Kim, and Lee [74] provide the rare mobility-condition figure: locating a target among five candidates and recognising its status took approximately 5 s, with 93% accuracy seated and 86% while walking.

Beyond-touch input is the other half of the answer to RQ2. Gong, Yang, and Irani [27] demonstrated one-handed continuous wrist-gesture input (WristWhirl) with recognition of 93.8% rising to 95.4%, directional marks averaging 483 ms and free-form shapes 1,436 ms — gross-motor input at microinteraction speed. Perrault et al. [52] moved input onto the watch strap for eyes-free command selection; Rey et al. [56] characterised bezel-to-bezel gestures on circular watches across stationary, mobile, and explicitly *encumbered* conditions and concluded they suit rapid eyes-free invocation. Yeo et al. [72] derived pressure, twist, and pan input from a stock watch IMU with no hardware modification. On the force channel, Darbar, Sen, and Samanta [19] instrumented the *sides of the watch casing* with force-sensitive resistors, eliminating screen occlusion entirely; Ren and Arif [55] established comfortable force ranges and showed in a Fitts' law study that force-tap outperforms plain tap specifically for *smaller* targets; Villarreal-Narvaez et al. [67] built a squeeze-gesture taxonomy of 82 classes and elicited 672 gestures from 32 participants, collapsing to 26 consensus classes; and Yi et al. [73] used squeezing of the device body as a shoulder-surfing-resistant private input channel. Kim, Son, and Oakley [32] round out the envelope with around-device selection on an unmodified watch: for target widths of 3, 6, and 9 mm, double-crossing achieved 2.18 bps at 0.88 s with 3.29% error, dwell 1.48 bps at 1.34 s with 0% error, and pinch 0.92 bps at 1.48 s with 7.52% error. The zero-error dwell condition is significant for our purposes: where the cost of a false actuation is high, trading throughput for accuracy is the correct exchange.

2.4 Covert and deniable interaction

Plausible deniability is well theorised in security and thinly evaluated in HCI. The strongest empirical precedent we could verify is Delusio [21], in which a particular facial expression unlocks a hidden secure area so that a coerced unlock yields an innocuous device; a 50-participant role-play study found that "suspects" accessed the hidden area largely without the coercer detecting it, and participants wanted the feature. Two properties of that work transfer directly to our framework: the design commitment that the *innocuous* state is what a coercer sees by default, and the methodological commitment to evaluating covertness adversarially through role-play rather than through accuracy metrics alone. Beyond this, we found no verifiable peer-reviewed evaluation of duress codes, disguised-icon safety applications, or stealth modes on wearables — a gap this framework's State 2 is formulated to address.

3 Proposed IxD Framework and State Architecture

3.1 Design commitments

The framework rests on four commitments, each traceable to the evidence above.

C1: Detection produces evidence, not verdicts. No wrist-sensing classifier reported in the literature approaches the specificity required for autonomous external escalation (§1.2). The system therefore never treats inference as sufficient grounds for contacting a responder. Inference is sufficient only to *ask*, and the asking must itself be cheap enough to be survivable at the frequency the classifier's error rate implies.

C2: The interaction budget shrinks as the state escalates. Under threat, available motor precision and gaze availability fall while urgency rises. The framework therefore inverts the usual confirmation pattern: the highest-consequence transitions require the *least* precise motor act, and the burden of interaction falls on *cancellation* rather than on *activation*.

C3: The covert presentation is the default, not a mode. Following [26, 12, 7, 21], the system's visible surface during active protection is indistinguishable from ordinary use. A "stealth mode" that the user must enable is a mode the coerced user cannot safely enable.

C4: Escalation to external responders is a consent event, not a threshold crossing. Following [8, 30, 3], irreversible consequential action requires either explicit user assent or the explicit, pre-configured, and auditable absence of a veto within a bounded window.

3.2 Sensor fusion matrix

Table 1 specifies the evidence vectors. We distinguish the three vectors the blueprint requires — physiological, environmental, behavioural — and for each signal we state the derivation, the discriminative claim it licenses, its documented failure mode, and the fusion level at which it enters, following Gravina et al.'s taxonomy [28]. The failure-mode column is load-bearing: a fusion architecture that does not encode what each channel gets wrong will weight the channels wrongly.

Table 1. Sensor fusion matrix. Fusion level per Gravina et al. [28]. All evidence is evaluated on-device; see §4.2.

Vector	Signal	Derivation / window	Discriminative claim	Documented failure mode	Fusion level
Physiological	Heart rate (PPG)	1 Hz, 30 s rolling median vs. personal 14-day	Sustained elevation unexplained by concurrent	Wrist PPG agrees with ECG for HR (r = .99) but motion artefact	Feature

		circadian baseline	motion intensity	dominates during vigorous movement [62]	
Physiological	HRV — SDNN, HF	120 s window, artefact-rejected	Withdrawal of parasympathetic tone consistent with acute arousal	E4-class wrist PPG: SDNN r = .91, HF r = .88 vs. ECG, but RMSSD shows >25% difference factor — the short-window index most wanted for acute stress is the least trustworthy [62]	Feature
Physiological	Electrodermal activity	4 Hz, tonic level + phasic response count	Sympathetic activation; the modality most closely correlated with stress level in real-world driving [29]	Confounded by ambient temperature, exertion, and hydration; no consumer watch ships a validated wrist EDA channel for this purpose	Feature
Physiological	Arousal-type disambiguation	Joint HR/EDA/motion model	Distinguishes exertion-driven from threat-driven arousal — required because arousal magnitude	Exercise, caffeine, febrile illness, panic disorder, and arrhythmia all produce	Decision
			alone does not predict motor degradation [49]	threat-like signatures	
Behavioural	Gross motion anomaly	50 Hz tri-axial accelerometer + gyroscope, 2 s windows	Struggle-like motion: punch, shove, slap, shake, discriminated against clap, wave, handshake, door-opening, typing	Single dominant-wrist configuration: 98.1% accuracy but specificity only 93.6% — bilateral sensing recovers 99.8% but is not a deployable assumption [64]	Feature
Behavioural	Gait and posture discontinuity	Cadence, stance variance, orientation	Abrupt cessation of gait, transition to prone, sustained immobility after high-magnitude impact	Real-world precision collapses to 0.37 despite 0.99 accuracy on class-imbalanced deployment data [39]; 4.7% sensitivity outside the ambulatory tuning population [1]	Feature
Behavioural	Interaction-pattern deviation	Device-use rhythm vs. personal model	Uncharacteristic absence of routine interaction during a normally active window; abandoned	Confounded by ordinary routine variation; very low specificity alone	Decision

			-device signature		
Environmental	Spatial isolation index	GPS/GNSS + on-device map priors + time-of-day	Wearer is in a low-traffic location at an atypical hour, off habitual route	Indoor and urban-canyon degradation; map priors encode socioeconomic bias about which areas are "unsafe" — see §5.4	Feature
Environmental	Route deviation	Comparison to learned habitual trajectories	Departure from the wearer's own historical routes, optionally from a shared itinerary	Legitimate novelty is indistinguishable from coerced deviation	Feature
Environmental	Acoustic event class	On-device classifier: raised voice, glass break, impact, scream-band energy	Corroborating evidence for an interpersonal event	Continuous audio classification is the highest-privacy-risk channel in the system; must never leave the device as audio in States 0–1 (§4.2)	Feature
Environmental	Ambient light / display state	Photodiode, screen state	Device concealed (pocket, sleeve) vs. exposed — governs whether visual signalling is even perceivable	Sleeve occlusion is indistinguishable from concealment	Feature
Fusion output	Threat evidence score	Weighted decision-level combination with per-channel reliability discount	Graded posterior over {no event, ambiguous event, probable event} with explicit uncertainty	Correlated channel failure during vigorous motion degrades multiple channels simultaneously — the architecture must detect and declare this rather than fusing corrupted inputs	Decision

Two properties of Table 1 deserve emphasis. First, no single channel is permitted to drive a transition. The physiological channels are the least trustworthy in exactly the conditions of interest (motion artefact during struggle degrades PPG precisely when struggle is what we wish to detect), so the architecture requires corroboration *across* vectors and applies a per-channel reliability discount conditioned on measured signal quality. Second, the architecture must be able to say "I cannot currently sense reliably." Correlated channel failure — vigorous motion corrupting PPG and HRV while also driving the motion classifier toward a positive — is the most dangerous failure mode in the system, because it produces high apparent confidence from a single underlying cause. Following the intelligibility argument [8, 37], this condition is surfaced to the user rather than silently absorbed.

3.3 The 4-State Transition Architecture

Table 2 specifies the architecture. Each state is defined by five properties: the evidence or act that licenses entry, the system's behaviour, the visible surface presented to an observer, the actions available to the wearer, and the transition rules with their dwell times. The structure is a ladder with asymmetric rungs: upward transitions from State 0 to State 1 are cheap and reversible; the transition from State 2 to State 3 is the only irreversible one and is the only transition that requires either explicit assent or an audited veto window.

Table 2. The 4-State Transition Architecture.

Property	State 0 — Ambient Baseline	State 1 — Micro-Verification	State 2 — Silent Beacon	State 3 — Critical Broadcast
Entry condition	Default. Entered on system start and on any downward transition.	Fusion score crosses the *ambiguity* threshold (corroboration required across ≥2 vectors), **or** wearer performs an initiation act (double casing squeeze, or bezel hold)	Wearer confirms in State 1, **or** State 1 veto window elapses with no response **and** the pre-configured no-response policy is set to escalate, **or** wearer performs the direct-to-beacon act	Wearer confirms in State 2, **or** State 2 evidence reaches the *probable event* threshold **and** the pre-configured dispatch policy authorises auto-escalation after the veto window
System behaviour	On-device sensing at low duty cycle; personal baselines updated; no data leaves the device; no session opened	Sensing at full rate; evidence buffer begins recording to encrypted on-device ring buffer (not transmitted); no contact with any external party	Encrypted location stream and optional audio stream to the wearer's pre-selected guardians; evidence buffer sealed and time-stamped; no audible or visible alarm; check-in cadence begins	Location and evidence transmitted to guardians *and*, per policy, to the emergency path; maximum-volume siren, display strobe, high-contrast alert screen; continuous broadcast until authenticated stand-down
Visible surface	Ordinary watch face	**No change to the visible surface.** The entire query is delivered vibrotactilely. If the display is already active, an unobtrusive glanceable badge is permitted, sized per §3.4	**Covert by construction:** the watch face is the wearer's ordinary face. Beacon status is legible only through a pre-agreed, deniable indicator the wearer chooses at setup (e.g. a specific complication colour state) and through a distinct low-amplitud	Deliberately maximally salient: high-contrast full-screen alert, siren, strobe. Visibility is the intervention

			e haptic heartbeat	
Wearer actions available	Initiate State 1 or State 2 directly; configure policy; review audit log	**Confirm** (escalate to 2): one gross-motor act. **Dismiss** (return to 0): one distinct gross-motor act. **Defer**: no action within the window	**Escalate** to 3; **stand down** to 0 with authentication; **extend** beacon duration; answer guardian check-in with a single haptic-acknowledged gesture	**Stand down** only with authentication (PIN or biometric), because this is the one control an adversary has a motive to operate
Transition up	to State 1 on evidence or initiation act	to State 2 on confirm, on policy-governed non-response, or on evidence reaching *probable*	to State 3 on wearer escalation, or on policy-authorised evidence threshold after veto window	Terminal
Transition down	—	to State 0 on dismiss, on evidence decay below threshold, or on window expiry under a no-escalate policy	to State 0 on authenticated stand-down or beacon expiry with guardian acknowledgement	to State 0 on authenticated stand-down only
Interaction budget	None	≤ 4 s to perceive and respond, matching the microinteraction envelope [4] and the 5 s median session [68]; veto window 15–30 s,	Zero obligatory interaction; check-ins acknowledged in ≤ 2 s or auto-acknowledged by continued motion	Zero interaction required to sustain; authenticated act required to stop
		user-configurable		
Governing design rationale	Data minimisation [34]; no sensing product leaves the device	Mixed-initiative asking rather than acting [30]; *scope services when in doubt* and *support efficient dismissal* [3]; relocates Karusala and Kumar's activation friction [31] to escalation	Covert-by-default against the UI-bound adversary [26, 21]; asynchronous resumable help channel [66]; avoids alarm aesthetics that trigger hypervigilance [13]	Consent-gated irreversibility [8]; minimum-data-set payload by analogy to the eCall regulatory precedent [25] and NG9-1-1 additional-data conventions [42]

3.4 Interaction and ergonomic modalities

RQ2 asks which modalities survive. Table 3 maps candidate modalities onto the motor requirement each imposes, the published performance envelope, and the states in which each is appropriate. The organising principle is a *substitution ladder*: as the state escalates and available precision falls, the required act moves from fingertip targeting toward whole-limb and whole-hand action.

Table 3. Low-dexterity modality ladder.

Modality	Motor requirement	Published envelope	Appropriate states	Principal failure mode
Fingertip touch target	Fine positional control, 4.2–9.6 mm depending on encumbrance	9.2 mm discrete / 9.6 mm serial one-handed thumb [50]; 4.2 mm minimum radius and 48.1% / 31–34% accuracy when encumbered and walking [43]; 80% error reduction when occlusion removed [71]	0 only, and never as the sole path in 1–3	Degrades exactly where needed; occluded by the acting finger

Dwell selection	Positional control but *no* discrete actuation	1.48 bps, 1.34 s, **0% error** around-device [32]	1 (as a secondary confirm)	Slow; requires sustained gaze
Bezel-to-bezel gesture	Coarse rotational sweep across the case edge	Characterised stationary, mobile, and **encumbered**; suited to rapid eyes-free invocation [56]	0, 1, 2	Requires the second hand
Casing squeeze (side force)	Whole-hand grip; force generation, not fine modulation	Side-mounted FSRs avoid screen occlusion entirely [19]; comfortable force ranges established and force-tap superior for small targets [55]; 26 consensus squeeze classes from 672 elicited gestures [67]; squeeze as a shoulder-surfing-resistant private channel [73]	**1, 2, 3 — the primary channel**	Requires the second hand; risk of accidental actuation during carrying
Wrist rotation / whirl	Gross forearm rotation, one-handed	93.8–95.4% recognition; directional marks 483 ms, free-form 1,436 ms [27]; wrist-rotation menu with tacton status, 93% seated / 86% walking [74]	**1, 2 — the primary one-handed channel**	Confusable with ordinary arm motion; needs a delimiting posture
Strap-surface input	Coarse touch off the display	Eyes-free command selection on the strap [52]	1, 2	Requires the second hand; strap must be instrumented
IMU-derived twist / pressure	Coarse manipulation of the whole device	Pressure, twist, and pan from a stock IMU, no hardware change [72]	1	Lower precision; susceptible to motion noise
Vibrotactile output (tacton)	None (perception only)	2 parameters ≈ 70% identification; 48% with 3; 81% with reduced values on one parameter [11]; 84 icons distinguishable via rhythm, discriminated by note length and unevenness [65]; wrist-array alert perception under multitasking [36]; localisation governed by body landmarks [14]	**All states — the primary output channel**	Capacity is small; must not be confused with ordinary notifications

Three specifications follow from the table.

Vibrotactile grammar. Because identification collapses when a third parameter is added [11] and scales through rhythm [65], the safety grammar uses **rhythm as the sole discriminative dimension** and reserves a distinct amplitude envelope purely to mark the message as safety-critical. Three patterns suffice for the architecture: an *interrogative* (two short pulses, ascending gap — "State 1 entered, respond"), a *confirmatory* (one long pulse — "State 2 active, beacon running"), and a *heartbeat* (a single low-amplitude pulse at a slow fixed period — "beacon still running", designed to be perceptible to the wearer and inaudible at conversational distance). This keeps the set at three within one parameter, well inside the 81% identification regime, and reserves no meaning to body location, which wrist form factors cannot reliably deliver [14]. The patterns must be assigned exclusively to the safety subsystem and never reused by notifications, because a pattern the wearer has learned to ignore is worse than no pattern.

Gross-motor confirmation. State 1 confirmation is a **double casing squeeze**; State 1 dismissal is a **single sustained wrist rotation** held past a threshold angle. This pairing is deliberate. The two acts are kinematically dissimilar, so a recognition error is unlikely to convert one into the other; both are gross-motor, avoiding the degraded fine-force channel [49]; the confirm act is the one more easily performed with a

restrained arm, while the dismiss act is one-handed and available when the other hand is occupied. Crucially, the *error asymmetry is designed*: a missed confirm results in the veto window elapsing and policy taking over, whereas a false confirm results in a silent beacon — a recoverable, non-public consequence — not a siren.

Threshold and debounce specification. Each gesture carries an actuation threshold tuned for *low false-positive rate rather than low latency*, on the reasoning that the cost asymmetry at State 1 favours the veto window over a fast but fragile trigger; this is the same exchange the zero-error dwell condition represents [32]. Every gesture is acknowledged by its own short tacton, so the wearer knows the act registered without looking — an indispensable property given elevated blink rates and gaze commitment to the threat [44].

3.5 Explicit interaction flows

Flow A — System-initiated, benign (the common case). Wearer walks an unfamiliar route at 23:40; the isolation index and route deviation both rise; HR is elevated but the motion signature is consistent with brisk walking, so the arousal-type model discounts the physiological evidence. The fusion score crosses the ambiguity threshold but not the probable-event threshold. → **State 1.** The interrogative tacton fires; the display does not change. The wearer performs a single sustained wrist rotation without breaking stride or looking at the watch; the dismissal tacton acknowledges. → **State 0.** Total elapsed interaction: under 3 s. No data has left the device; no contact has been made; the event is written to the local audit log and used to adjust the personal baseline. This flow is the one that must be survivable many times per week, and it is the reason State 1 exists.

Flow B — System-initiated, genuine, wearer unable to respond. Wearer is grabbed from behind. The motion classifier registers struggle-like motion; acoustic classification registers a raised voice; HR rises sharply while HRV-derived evidence is discounted for motion artefact. → **State 1** with the interrogative tacton. The wearer does not respond — both hands are occupied. The 20 s veto window elapses. Because the wearer's configured no-response policy is *escalate*, the system enters **State 2**: encrypted location streaming to two guardians, audio capture to the sealed evidence buffer, the ordinary watch face unchanged, the heartbeat tacton confirming to the wearer alone that help has been signalled. Evidence continues to accumulate past the probable-event threshold; the dispatch policy authorises escalation; a second veto window elapses. → **State 3:** siren, strobe, high-contrast alert, minimum-data-set payload to the emergency path. Note that the wearer performed *no* motor act at any point, and that each escalation was authorised in advance by a policy the wearer set while calm.

Flow C — User-initiated, covert, coercive-control context. Wearer recognises a coercive situation developing in a domestic setting where the aggressor has physical access to the watch. The wearer performs a double casing squeeze with the device inside a sleeve — no gaze, no display change. → **State 2 directly**, bypassing State 1. Guardians receive location and a message the wearer pre-authored while calm. The watch face is unchanged; the pre-agreed deniable indicator reads as an ordinary complication state. If the aggressor takes the watch, the beacon continues and stand-down requires authentication the aggressor does not hold. If the aggressor demands the wearer stop it, the wearer can perform an authenticated stand-down that in fact enters a *decoy* stand-down — the design question of whether to offer this, and the risk it carries if discovered, is exactly the question Delusio's adversarial role-play method [21] exists to answer, and we treat it as an empirical question for §5 rather than a settled design choice.

Flow D — Recovery from a false State 3. Siren active in error. The wearer authenticates and stands down; the system transmits a structured cancellation to every party contacted, including the emergency path, rather than merely going silent — the responder-burden literature [58, 41] indicates that an uncancelled false alert consumes more response capacity than the initial alert. The event is logged with its full evidence trace, and the wearer is offered a one-tap adjustment of the policy that produced it. Following Kocielnik et al. [33], the system's error profile is stated in precision/recall terms at setup, so that the false alarm is a disappointment rather than a betrayal.

4 Ethical Design and Privacy Architecture

4.1 Continuous sensing against user agency

The framework's premise — that protection begins before the wearer acts — entails continuous sensing of the body, its motion, its location, and its acoustic surroundings. This is the most invasive sensing envelope in consumer computing, and it is assembled in service of a population disproportionately likely to be surveilled. Langheinrich's ubicomp privacy principles [34] give the structure we adopt: notice, choice and consent, proximity and locality, anonymity and pseudonymity, security, and access and recourse. Nissenbaum's contextual integrity [47] supplies the analytic move that matters most: the question is never whether biometric or location data is "sensitive" in the abstract but whether a particular *flow* conforms to the norms of the context. A location flow to a dispatcher during an assault is appropriate; the identical flow to an employer, an insurer, or an intimate partner is a violation, even when the recipient claims a protective motive — which, per Bellini [7], is exactly the claim abusive partners make about GPS tracking.

Barkhuus and Dey [6] found users accept reduced control when usefulness is high but remain sensitive to its loss, which we operationalise as a hard constraint: every proactive capability in the framework has an off switch, the off switch is reachable without authentication (so that it cannot be used

against a coerced user as a lock), and the *fact* of it being off is not broadcast to guardians unless the wearer chose that at configuration. Consolvo et al. [17] found that location-disclosure decisions depend on who is asking and why, and that people frequently prefer all-or-nothing disclosure over blurring; we therefore scope disclosure by *recipient and state* rather than offering a precision slider that users do not want and cannot reason about.

4.2 On-device evaluation versus transmission

The architecture draws a hard line at the State 1/State 2 boundary: **in States 0 and 1, no sensor product of any kind leaves the device.** All fusion, all baseline modelling, and all threshold evaluation are local. Audio in particular is classified on-device into event categories and the waveform is discarded; the sealed evidence buffer that begins in State 1 is written to encrypted local storage and is transmitted only on entry to State 2, when the wearer has either consented or allowed a pre-authorised veto window to elapse. This is not merely a privacy nicety. It is the property that makes the system's continuous sensing defensible at all, and it is the property that a cloud-inference architecture cannot offer, because in such an architecture the decision to transmit necessarily precedes the decision that transmission is warranted.

Personalisation is the obvious tension: per-wearer baselines are what make the ambiguity threshold meaningful, and per-population model improvement is what would raise specificity toward the levels §1.2 shows are needed. Federated averaging [40] is the relevant mechanism — model updates rather than raw streams — and we specify it with the honest caveat that federated learning reduces but does not eliminate inference risk, and that a safety population is precisely the population for which residual risk is least acceptable. Where a choice must be made between a measurable specificity gain and a raw-data collection requirement, the framework's commitment is to forgo the gain and disclose the resulting error rate, per [33].

4.3 The adversary who is already authenticated

Freed et al.'s UI-bound adversary [26] and Chatterjee et al.'s dual-use spyware ecosystem [12] establish that the threat model cannot assume an external attacker. Three architectural consequences follow.

First, **the covert state is the default** (C3). A protective state that is visually distinguishable from ordinary use is a state whose existence an abuser can observe, punish, and forbid. This is why State 2's visible surface is specified as *unchanged* rather than as *subtle*.

Second, **stand-down is the authenticated control, and activation is not.** The usual arrangement — authentication to activate, a tap to cancel — is exactly inverted from the threat model. The party with a motive to silence the system is the aggressor; the party with a motive to start it is the wearer.

Third, **guardian configuration is itself an attack surface.** An abuser with device access can add themselves as a guardian, in which case the beacon streams the wearer's location to the aggressor. The framework therefore requires that guardian *additions* take effect only after a delay with out-of-band notification to existing guardians, that the guardian list be viewable in a form that survives coercion (the audit log is append-only and its deletion is itself logged), and that the system never display a guardian's read receipts in a way that reveals to a co-present observer who is watching. Matthews et al. [38] establish that needs invert across the control, escape, and life-apart phases, so the framework additionally requires that the entire configuration be exportable and re-establishable on a new device without contacting the old one — a survivor replacing a compromised watch should not have to authenticate through the compromised account.

4.4 False positives, alert fatigue, and responder scarcity

The framework treats the false-positive budget as a first-class design resource allocated across three distinct costs, because they are bounded by different quantities.

The wearer's attention. Every State 1 interrogation costs a microinteraction. At a median session length of 5 s and with only 9.4% of ordinary notifications producing any interaction at all [68], a safety tacton that fires several times a day will be learned as ignorable — the desensitisation mechanism Cvach [18] and Drew et al. [23] document, where 88.8% false-positive rates and 187 audible alarms per bed-day produce non-response. The design target is therefore stated explicitly: **State 1 entries should be rare enough to remain salient** — our recommendation is a configurable ceiling enforced by an adaptive threshold, with the ceiling defaulting to a small number per week rather than per day, and with the enforcement mechanism visible in the audit log.

The wearer's social standing. Karusala and Kumar [31] found that fear of embarrassing false alarms was a first-order deterrent to panic-button use. State 2's covertness is what decouples a false positive from a social cost: a spurious silent beacon costs the wearer a message to a guardian, not a public scene.

Responder capacity. State 3 is the only state that consumes a shared resource, and the burglar-alarm precedent [58] shows how automated systems overwhelm the responders they recruit, while Ming et al. [41] show that the institutional responder side, not the reporting interface, determines whether an alert yields help. The framework therefore (a) defaults the emergency-path escalation to *off*, requiring explicit opt-in per the mixed-initiative expected-utility argument [30]; (b) transmits a minimum data set with an explicit confidence indication and the evidence basis, following the eCall

minimum-data-set model [25] and NG9-1-1 additional-data conventions [42]; and (c) mandates structured cancellation (Flow D). We note Wickens et al.'s finding [69] that a 45% false-alert rate need not produce cry-wolf behaviour where operators retain situational awareness — which supports telling responders *why* the system escalated rather than merely that it did.

Finally, we record a tension the framework does not resolve. Chordia et al. [15, 16] show that safety technology can increase anxiety without increasing safety, and that reducing dysfunctional fear is itself a legitimate design goal. A device that continuously evaluates one's surroundings for threat may cultivate the vigilance it purports to relieve. Our State 0 design commitment — no visible ambient threat signalling, ever, and no "risk score" surfaced to the wearer — is a direct response, but whether the framework as a whole reduces or amplifies fear of crime is an empirical question we specify as a measured outcome in §5 rather than assert.

5 Evaluation Protocol and Methodology

The framework makes claims of three kinds — that low-dexterity inputs remain operable under stress, that the state architecture achieves an acceptable error trade-off, and that State 2 is genuinely covert against an attentive observer — and each requires a different method. We specify a three-study programme. Throughout, we distinguish claims testable in a laboratory from claims that are not, and we state the latter as limitations rather than deferring them.

5.1 Study 1: Low-dexterity input under induced stress and cognitive load

Design. Within-subjects, 2 (stress: baseline vs. induced) × 4 (modality: fingertip touch on a 9 mm target, dwell selection, double casing squeeze, sustained wrist rotation), with modality order counterbalanced by Latin square and stress blocked with a washout period. Target N = 32, powered for a medium within-subjects effect on the primary measure; participants screened for cardiac conditions and for recent experience of violence, with the latter an exclusion on ethical rather than methodological grounds.

Stress and load induction. A laboratory cannot ethically reproduce the threat-type stressor that produces the ~130% steadiness decrement [49], and we do not claim otherwise. We therefore induce along two dimensions and report them separately. *Socially evaluative stress* uses the Trier Social Stress Test, as validated by Sarsenbayeva et al. [59] for mobile-interaction measurement, with the honest expectation that its effects on touch will be small (their offset penalty was roughly 4 px). *Physical and cognitive load* combines encumbrance and walking following Ng et al.'s protocol [43] — carrying bags while walking a prescribed route — with a concurrent auditory n-back task to load working memory. Induction is validated by HF-HRV and STAI [59], plus a manipulation-check NASA-TLX. Reporting the two induction dimensions separately is essential: conflating them is how the literature acquired its overconfident folklore about stress and touch.

Measures. *Time-to-Acknowledge (TTA)*, operationalised as the interval from tacton onset to a recognised, correctly classified gesture completion, with the microinteraction envelope of 4 s [4] as the a priori criterion. *Actuation success rate* per modality. *Cross-gesture confusion rate* — the proportion of confirm acts classified as dismiss and vice versa — which we treat as the safety-critical measure, since it is the one that inverts the wearer's intent. *Tacton identification accuracy* for the three-pattern grammar, both isolated and while walking, benchmarked against the 81% figure for reduced-parameter tactons [11] and the 93%/86% seated/walking figures for tacton status recognition [74]. *NASA-TLX* subscale profile per modality. *SUS* for the end-to-end State 0→1→0 interaction.

Hypotheses. H1: TTA for squeeze and wrist rotation remains within the 4 s envelope under both induction conditions, while fingertip touch on a 9 mm target does not under combined encumbrance and load, consistent with the 31–48% accuracy figures in [43]. H2: cross-gesture confusion is lower for the kinematically dissimilar squeeze/rotation pair than for two touch targets on the same display. H3: NASA-TLX mental-demand loading is lower for the tacton-only output channel than for a glanceable visual badge under the load condition.

5.2 Study 2: Error behaviour of the state architecture in situ

Specificity claims cannot be established in a laboratory, because the denominator — the wearer's ordinary life — is what the laboratory removes. Study 2 is therefore a multi-week in-the-wild deployment measuring the architecture's *false-positive* behaviour only, with State 2 and State 3 disabled at the transport layer so that no external party is ever contacted.

Design. N = 40 participants, 4 weeks, instrumented watch running the full State 0/State 1 logic. Every State 1 entry is logged with its full evidence trace and the wearer's response. Participants complete an experience-sampling item after each State 1 entry, and a weekly diary.

Measures. *False-positive rate* expressed per wearer-day and per decision window, with the 3.5×10^{-5} per-window figure from §1.2 as the reference target for what autonomous escalation would require. *Response distribution* across dismiss, confirm, and non-response — the non-response proportion is the critical unknown, because the architecture's no-response policy is what converts inattention into escalation. *Salience decay*: TTA and dismissal rate as a function of cumulative State 1 exposure, testing directly for the desensitisation documented in [23, 18]. *Fear-of-crime measures* administered at entry and exit, following Chordia et

al. [16], to test whether the framework reduces or amplifies dysfunctional fear. *Detection sensitivity is deliberately not claimed*; establishing it would require genuine threat events, and we treat that as outside what any ethical study design can deliver, noting that the field's substitute — induced-event protocols [9, 1] — is exactly what failed to generalise.

5.3 Study 3: Adversarial covertness evaluation

Covertness is a property relative to an observer and cannot be measured by self-report. We adopt Delusio's role-play method [21], in which one participant plays a coercer and another operates the covert feature, and the dependent measure is whether the coercer *detects* the operation.

Design. Dyadic role-play, N = 25 dyads, scripted scenarios of increasing observer attentiveness: an inattentive co-present observer, an attentive observer, and an observer explicitly briefed that a concealed safety feature may exist and instructed to find it. Dyads are composed of unacquainted participants, and no scenario depicts violence; the manipulation is attention, not threat.

Measures. *Detection rate* per attentiveness level, which is the primary outcome and which we expect to degrade sharply in the briefed condition — a result the framework must be able to absorb honestly, since a covert feature that is covert only against the uninformed protects only against an adversary who has not read the product documentation. *Time-to-detection*. *Operator confidence* versus actual detection, whose calibration gap is itself a safety hazard. *Deniability of the stand-down decoy* (Flow C), evaluated separately, with a pre-registered decision rule: if detection of the decoy exceeds a threshold in the briefed condition, the feature is not shipped, because a detected decoy converts a coercive situation into a punitive one.

Ethical governance. All three studies require review by an institutional ethics board with domain expertise in interpersonal violence, a safeguarding protocol with on-call support, participant-initiated termination without explanation at any point, and — for Study 3 — debriefing that makes explicit the fictional nature of the coercive frame. Recruitment does not target survivors; the framework's survivor-relevant claims are developed with advocacy organisations as consultants rather than with survivors as experimental participants, following the care-infrastructure and trauma-informed guidance of [66, 13].

5.4 Threats to validity

We record four. *Construct validity of stress induction*: the TSST is a socially evaluative stressor, not a threat stressor, and the gap between its effects [59] and those in the police-shooting paradigm [44, 46] is the single largest inferential weakness in any laboratory study of this kind. *Population validity*: fall-detection history [1] should make any wearable-safety researcher assume that a model validated on ambulatory young adults will fail for disabled, older, and differently-moving wearers, and Study 1 should be replicated with wheelchair users and with participants whose baseline motor variability is elevated. *Ecological validity of role-play covertness*: a briefed confederate is a weak proxy for an intimate with years of observational familiarity with the wearer's habits. *Bias in the environmental vector*: the spatial isolation index encodes map priors about which locations are dangerous, and those priors reflect the reporting patterns of policing and crowd-sourced crime data — precisely the data whose deployment Chordia et al. [15] show to amplify anxiety and social injustice. An audit of the isolation index for demographic and geographic bias is a required component of the evaluation, not an optional extension.

6 Design Guidelines and Conclusion

6.1 Five heuristics for safety-critical wearable interfaces

H1. Escalate on evidence; act only on consent. Inference licenses a question, not an action. Where a system's action is irreversible or consumes a shared responder resource, the transition must be authorised by explicit assent or by the elapse of a veto window under a policy the wearer configured while calm. This is Horvitz's expected-utility threshold [30] and Bellotti and Edwards' accountability requirement [8] applied to a case where the cost of acting wrongly and the cost of failing to act are both severe and asymmetric. The corollary is architectural: build the cheap question before the expensive action, because a system with no State 1 will be forced to choose between a classifier it cannot trust and a button the wearer cannot press.

H2. Design the confirmation, not the alarm. The hard interaction problem in a safety system is not signalling an emergency; it is the seconds in which an ambiguous situation is resolved. Budget that interaction against the microinteraction envelope — under 4 s [4], against a 5 s median session [68] — deliver it in a channel the threat is not already occupying, and place the burden of action on *cancellation* rather than activation, so that the wearer who cannot act is not thereby unprotected.

H3. Degrade the input channel, not the intent. As stress and encumbrance rise, substitute gross-motor acts for fine ones rather than simplifying the goal. Threat-type stressors degrade fine force regulation while leaving force generation intact [49, 48], so a squeeze, a wrist rotation, or a bezel sweep remains available where a 9 mm target does not [43, 50, 27, 56, 19]. Pair the confirm and cancel acts so that they are kinematically *dissimilar*, and design the recognition thresholds so that the residual errors fall on the recoverable side.

H4. Make the deniable state the default state. Where the adversary may be co-present and device-authenticated [26, 12, 7], a protective mode that looks different from ordinary use is

a mode the adversary can observe and forbid. Invert the authentication asymmetry: activation requires none, stand-down requires authentication, because the party with a motive to stop the system is not the wearer. Then evaluate covertness adversarially [21], including against an observer who knows the feature exists.

H5. Budget false positives against the wearer's attention, the wearer's social standing, and the responders' capacity — separately. These three costs are bounded by different quantities and are paid to different parties, so a single "accuracy" figure cannot govern them. Compute the per-window false-positive rate that the deployment duty cycle implies before claiming a classifier is deployable; state the resulting error profile to the wearer in precision/recall terms at setup [33]; make a false positive in the covert state socially costless; and mandate structured cancellation on the one path that consumes a shared resource [58, 41].

6.2 Theoretical contributions

The framework makes three contributions to wearable HCI beyond the specific design. First, it reframes threat recognition as a *state negotiation* rather than a classification problem, which relocates the design effort from the detector to the transition interface — a move we believe generalises to any wearable inference whose action is consequential and whose classifier is imperfect, including medical alerting and driver monitoring. Second, it articulates an *inverted interaction budget*: a design regime in which the required motor precision decreases as system state severity increases, which is the opposite of the confirmation-dialogue convention and which is derivable from the dissociation between fine-force degradation and preserved force generation under threat. Third, it positions *deniability as a default property of the visible surface* rather than as a mode, and in doing so brings the intimate-partner-violence threat model — where the adversary is authenticated and co-present — into the design core of a mainstream wearable capability rather than treating it as a specialist accommodation.

6.3 Limitations and future work

The framework is specified and argued but not yet implemented or evaluated; §5 is a protocol, not a result. Three limitations are structural rather than incidental. The specificity required for the architecture's upper transitions exceeds what any published wrist-sensing classifier demonstrates, and while State 1 is designed precisely to absorb that gap, the gap's size is an empirical unknown until Study 2 is run. The covertness claim is evaluable only against briefed confederates, not against intimates with years of observational familiarity. And the framework inherits the field's central asymmetry: its false-positive behaviour is measurable and its sensitivity is not, because the events it exists to detect cannot ethically be produced. We regard the appropriate response to that asymmetry not as optimism about detection but as architectural humility — which is what the four-state ladder, with its cheap questions and its consent-gated escalations, is for.

References

[1] Libak Abou, Alexander Fliflet, Lina Hawari, Peter Presti, Jacob J. Sosnoff, Harshal P. Mahajan, Mikaela L. Frechette, and Laura A. Rice. 2022. Sensitivity of Apple Watch fall detection feature among wheelchair users. *Assistive Technology* 34, 5, 619–625. DOI: 10.1080/10400435.2021.1923087

[2] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a Better Understanding of Context and Context-Awareness. In *Handheld and Ubiquitous Computing (HUC '99)*, LNCS 1707. Springer, 304–307. DOI: 10.1007/3-540-48157-5_29

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–13. DOI: 10.1145/3290605.3300233

[4] Daniel L. Ashbrook. 2010. *Enabling Mobile Microinteractions*. Ph.D. Dissertation. Georgia Institute of Technology.

[5] Ling Bao and Stephen S. Intille. 2004. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive Computing*, LNCS 3001. Springer, 1–17. DOI: 10.1007/978-3-540-24646-6_1

[6] Louise Barkhuus and Anind Dey. 2003. Is Context-Aware Computing Taking Control away from the User? Three Levels of Interactivity Examined. In *UbiComp 2003: Ubiquitous Computing*, LNCS 2864. Springer, 149–156. DOI: 10.1007/978-3-540-39653-6_12

[7] Rosanna Bellini. 2024. Abusive Partner Perspectives on Technology Abuse: Implications for Community-based Violence Prevention. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1. DOI: 10.1145/3637292

[8] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human-Computer Interaction* 16, 2–4, 193–212. DOI: 10.1207/S15327051HCI16234_05

[9] Bruce Brew, Steven G. Faux, and Elizabeth Blanchard. 2022. Effectiveness of a Smartwatch App in Detecting Induced Falls: Observational Study. *JMIR Formative Research* 6, 3, e30121. DOI: 10.2196/30121

[10] Stephen A. Brewster and Lorna M. Brown. 2004. Tactons: Structured Tactile Messages for Non-Visual Information Display. In *Proceedings of the 5th Australasian User Interface Conference (AUIC 2004)*, CRPIT Vol. 28. Australian Computer Society, 15–23.

[11] Lorna M. Brown, Stephen A. Brewster, and Helen C. Purchase. 2006. Multidimensional tactons for non-visual information presentation in mobile devices. In *Proceedings of MobileHCI '06*. ACM, 231–238. DOI: 10.1145/1152215.1152265

[12] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. 2018. The Spyware Used in Intimate Partner Violence. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 441–458. DOI: 10.1109/SP.2018.00061

[13] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A. Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. DOI: 10.1145/3491102.3517475

[14] Roger W. Cholewiak and Amy A. Collins. 2003. Vibrotactile localization on the arm: Effects of place, space, and age. *Perception & Psychophysics* 65, 7, 1058–1077. DOI: 10.3758/BF03194834

[15] Ishita Chordia, Lena-Phuong Tran, Tala June Tayebi, Emily Parrish, Sheena Erete, Jason Yip, and Alexis Hiniker. 2023. Deceptive Design Patterns in Safety Technologies: A Case Study of the Citizen App. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. DOI: 10.1145/3544548.3581258

[16] Ishita Chordia, JaeWon Kim, Zhuqing Liu, Hayley Park, Lucas Garrett, Sheena Erete, Christopher A. Le Dantec, Jason Yip, and Alexis Hiniker. 2024. Tuning into the World: Designing Community Safety Technologies to Reduce Dysfunctional Fear of Crime. In *Proceedings of the ACM Designing Interactive Systems Conference (DIS '24)*. ACM. DOI: 10.1145/3643834.3661578

[17] Sunny Consolvo, Ian E. Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. 2005. Location Disclosure to Social Relations: Why, When, & What People Want to Share. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, 81–90. DOI: 10.1145/1054972.1054985

[18] Maria Cvach. 2012. Monitor alarm fatigue: an integrative review. *Biomedical Instrumentation & Technology* 46, 4, 268–277. DOI: 10.2345/0899-8205-46.4.268

[19] Rajkumar Darbar, Prasanta Kr. Sen, and Debasis Samanta. 2016. PressTact: Side Pressure-Based Input for Smartwatch Interaction. In *Proceedings of the 2016 CHI Conference Extended Abstracts (CHI EA '16)*. ACM. DOI: 10.1145/2851581.2892436

[20] Anind K. Dey. 2001. Understanding and Using Context. *Personal and Ubiquitous Computing* 5, 1, 4–7. DOI: 10.1007/s007790170019

[21] Felix Dietz, Lukas Mecke, Daniel Riesner, and Florian Alt. 2024. Delusio: Plausible Deniability For Face Recognition. *Proceedings of the ACM on Human-Computer Interaction* 8, MHCI. DOI: 10.1145/3676494

[22] Gregory R. Dirkin. 1983. Cognitive Tunneling: Use of Visual Information under Stress. *Perceptual and Motor Skills* 56, 1, 191–198. DOI: 10.2466/pms.1983.56.1.191

[23] Barbara J. Drew, Patricia Harris, Jessica K. Zègre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. 2014. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PLoS ONE* 9, 10, e110274. DOI: 10.1371/journal.pone.0110274

[24] James A. Easterbrook. 1959. The effect of emotion on cue utilization and the organization of behavior. *Psychological Review* 66, 3, 183–201. DOI: 10.1037/h0047707

[25] European Parliament and Council of the European Union. 2015. Regulation (EU) 2015/758 concerning type-approval requirements for the deployment of the eCall in-vehicle system based on the 112 service. *Official Journal of the European Union* L 123 (19 May 2015), 77.

[26] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 1–13. DOI: 10.1145/3173574.3174241

[27] Jun Gong, Xing-Dong Yang, and Pourang Irani. 2016. WristWhirl: One-handed Continuous Smartwatch Input using Wrist Gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, 861–872. DOI: 10.1145/2984511.2984563

[28] Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. 2017. Multi-Sensor Fusion in Body Sensor Networks: State-of-the-Art and Research Challenges. *Information Fusion* 35, 68–80. DOI: 10.1016/j.inffus.2016.09.005

[29] Jennifer A. Healey and Rosalind W. Picard. 2005. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent*

Transportation Systems 6, 2, 156–166. DOI: 10.1109/TITS.2005.848368

[30] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, 159–166. DOI: 10.1145/302979.303030

[31] Naveena Karusala and Neha Kumar. 2017. Women's Safety in Public Spaces: Examining the Efficacy of Panic Buttons in New Delhi. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 3340–3351. DOI: 10.1145/3025453.3025532

[32] Jiwan Kim, Jiwan Son, and Ian Oakley. 2025. Cross, Dwell, or Pinch: Designing and Evaluating Around-Device Selection Methods for Unmodified Smartwatches. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM. DOI: 10.1145/3706598.3714308

[33] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–14. DOI: 10.1145/3290605.3300641

[34] Marc Langheinrich. 2001. Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems. In *UbiComp 2001: Ubiquitous Computing*, LNCS 2201. Springer, 273–291. DOI: 10.1007/3-540-45427-6_23

[35] Oscar D. Lara and Miguel A. Labrador. 2013. A Survey on Human Activity Recognition Using Wearable Sensors. *IEEE Communications Surveys & Tutorials* 15, 3, 1192–1209. DOI: 10.1109/SURV.2012.110112.00192

[36] Seungyon Claire Lee and Thad Starner. 2010. BuzzWear: alert perception in wearable tactile displays on the wrist. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 433–442. DOI: 10.1145/1753326.1753392

[37] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, 2119–2128. DOI: 10.1145/1518701.1519023

[38] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. 2017. Stories from Survivors: Privacy & Security Practices when Coping with Intimate Partner Abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 2189–2201. DOI: 10.1145/3025453.3025875

[39] Taylor R. Mauldin, Marc E. Canby, Vangelis Metsis, Anne H. H. Ngu, and Coralys Cubero Rivera. 2018. SmartFall: A Smartwatch-Based Fall Detection System Using Deep Learning. *Sensors* 18, 10, 3363. DOI: 10.3390/s18103363

[40] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 54, 1273–1282.

[41] Shufan Ming, Ryan Mayfield, Haocong Cheng, Ke-Rou Wang, and Yun Huang. 2021. Examining Interactions Between Community Members and University Safety Organizations through Community-Sourced Risk Systems. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1. DOI: 10.1145/3449111

[42] National Emergency Number Association. 2021. *NENA i3 Standard for Next Generation 9-1-1*. NENA-STA-010.3b-2021 (ANSI-approved 7 October 2021).

[43] Alexander Ng, Stephen A. Brewster, and John H. Williamson. 2014. Investigating the effects of encumbrance on one- and two-handed interactions with mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 1981–1990. DOI: 10.1145/2556288.2557312

[44] Arne Nieuwenhuys and Raôul R. D. Oudejans. 2010. Effects of anxiety on handgun shooting behavior of police officers: a pilot study. *Anxiety, Stress & Coping* 23, 2, 225–233. DOI: 10.1080/10615800902977494

[45] Arne Nieuwenhuys and Raôul R. D. Oudejans. 2012. Anxiety and perceptual-motor performance: toward an integrated model of concepts, mechanisms, and processes. *Psychological Research* 76, 6, 747–759. DOI: 10.1007/s00426-011-0384-x

[46] Arne Nieuwenhuys, Rouwen Cañal-Bruland, and Raôul R. D. Oudejans. 2012. Effects of Threat on Police Officers' Shooting Behavior: Anxiety, Action Specificity, and Affective Influences on Perception. *Applied Cognitive Psychology* 26, 4, 608–615. DOI: 10.1002/acp.2838

[47] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1, 119.

[48] J. Timothy Noteboom, Kerry R. Barnholt, and Roger M. Enoka. 2001. Activation of the arousal response and impairment of performance increase with anxiety and stressor intensity. *Journal of Applied Physiology* 91, 5, 2093–2101. DOI: 10.1152/jappl.2001.91.5.2093

[49] J. Timothy Noteboom, Monika Fleshner, and Roger M. Enoka. 2001. Activation of the arousal response can impair

performance on a simple motor task. *Journal of Applied Physiology* 91, 2, 821–831. DOI: 10.1152/jappl.2001.91.2.821

[50] Pekka Parhi, Amy K. Karlson, and Benjamin B. Bederson. 2006. Target Size Study for One-Handed Thumb Use on Small Touchscreen Devices. In *Proceedings of MobileHCI '06*. ACM, 203–210. DOI: 10.1145/1152215.1152260

[51] Sangkeun Park, Sujin Kwon, and Uichin Lee. 2018. CampusWatch: Exploring Communitysourced Patrolling with Pervasive Mobile Technology. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW. DOI: 10.1145/3274403

[52] Simon T. Perrault, Eric Lecolinet, James Eagan, and Yves Guiard. 2013. Watchit: simple gestures and eyes-free interaction for wristwatches and bracelets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 1451–1460. DOI: 10.1145/2470654.2466192

[53] Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. 2016. Smartwatch in vivo. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 5456–5469. DOI: 10.1145/2858036.2858522

[54] Esa M. Rantanen and Joseph H. Goldberg. 1999. The effect of mental workload on the visual field size and shape. *Ergonomics* 42, 6, 816–834. DOI: 10.1080/001401399185315

[55] Yuan Ren and Ahmed Sabbir Arif. 2023. Investigating a Force-Based Selection Method for Smartwatches in a 1D Fitts' Law Study and Two New Character-Level Keyboards. In *Proceedings of the 17th International Conference on Tangible, Embedded, and Embodied Interaction (TEI '23)*. ACM. DOI: 10.1145/3569009.3572741

[56] Bradley Rey, Kening Zhu, Simon T. Perrault, Sandra Bardot, Ali Neshati, and Pourang Irani. 2022. Understanding and Adapting Bezel-to-Bezel Interactions for Circular Smartwatches in Mobile and Encumbered Scenarios. *Proceedings of the ACM on Human-Computer Interaction* 6, ISS. DOI: 10.1145/3546736

[57] Daniel Salber, Anind K. Dey, and Gregory D. Abowd. 1999. The Context Toolkit: Aiding the Development of Context-Enabled Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, 434–441. DOI: 10.1145/302979.303126

[58] Rana Sampson. 2011. *False Burglar Alarms*, 2nd ed. Problem-Specific Guides Series No. 5. U.S. Department of Justice, Office of Community Oriented Policing Services, Washington, DC.

[59] Zhanna Sarsenbayeva, Niels van Berkel, Danula Hettiachchi, Weiwei Jiang, Tilman Dingler, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2019. Measuring the Effects of Stress on Mobile Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1, Article 24. DOI: 10.1145/3314411

[60] Bill N. Schilit, Norman Adams, and Roy Want. 1994. Context-Aware Computing Applications. In *Proceedings of the First Workshop on Mobile Computing Systems and Applications (WMCSA)*. IEEE, 85–90. DOI: 10.1109/WMCSA.1994.16

[61] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. ACM, 400–408. DOI: 10.1145/3242969.3242985

[62] Angela A. T. Schuurmans, Peter de Looff, Karin S. Nijhof, Catarina Rosada, Ron H. J. Scholte, Arne Popma, and Roy Otten. 2020. Validity of the Empatica E4 Wristband to Measure Heart Rate Variability (HRV) Parameters: a Comparison to Electrocardiography (ECG). *Journal of Medical Systems* 44, 11, Article 190. DOI: 10.1007/s10916-020-01648-w

[63] Mark A. Staal. 2004. *Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework*. NASA/TM-2004-212824. NASA Ames Research Center, Moffett Field, CA.

[64] Franck Tchuente, Natalie Baddour, and Edward D. Lemaire. 2020. Classification of Aggressive Movements Using Smartwatches. *Sensors* 20, 21, 6377. DOI: 10.3390/s20216377

[65] David Ternes and Karon E. MacLean. 2008. Designing Large Sets of Haptic Icons with Rhythm. In *Haptics: Perception, Devices and Scenarios (EuroHaptics 2008)*, LNCS 5024. Springer, 199–208. DOI: 10.1007/978-3-540-69057-3_24

[66] Emily Tseng, Mehrnaz Sabet, Rosanna Bellini, Harkiran Kaur Sodhi, Thomas Ristenpart, and Nicola Dell. 2022. Care Infrastructures for Digital Security in Intimate Partner Violence. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. DOI: 10.1145/3491102.3502038

[67] Santiago Villarreal-Narvaez, Arthur Sluÿters, Jean Vanderdonckt, and Efrem Mbaki Luzayisu. 2022. Theoretically-Defined vs. User-Defined Squeeze Gestures. *Proceedings of the ACM on Human-Computer Interaction* 6, ISS, 73–102. DOI: 10.1145/3567805

[68] Aku Visuri, Zhanna Sarsenbayeva, Niels van Berkel, Jorge Goncalves, Reza Rawassizadeh, Vassilis Kostakos, and Denzil Ferreira. 2017. Quantifying Sources and Types of Smartwatch Usage Sessions. In *Proceedings of the 2017 CHI*

Conference on Human Factors in Computing Systems (CHI '17). ACM. DOI: 10.1145/3025453.3025817

[69] Christopher D. Wickens, Stephen Rice, David Keller, Shaun Hutchins, Jamie Hughes, and Krisstal Clayton. 2009. False Alerts in Air Traffic Control Conflict Alerting System: Is There a "Cry Wolf" Effect? *Human Factors* 51, 4, 446–462. DOI: 10.1177/0018720809344720

[70] Meghan Woo and Olivia Bacon. 2020. Alarm Fatigue. In *Making Healthcare Safer III: A Critical Analysis of Existing and Emerging Patient Safety Practices*. Agency for Healthcare Research and Quality, Rockville, MD.

[71] Haijun Xia, Tovi Grossman, and George Fitzmaurice. 2015. NanoStylus: Enhancing Input on Ultra-Small Displays with a Finger-Mounted Stylus. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology (UIST '15)*. ACM, 447–456. DOI: 10.1145/2807442.2807500

[72] Hui-Shyong Yeo, Juyoung Lee, Andrea Bianchi, and Aaron Quigley. 2016. WatchMI: pressure touch, twist and pan gesture input on unmodified smartwatches. In *Proceedings of MobileHCI '16*. ACM, 394–399. DOI: 10.1145/2935334.2935375

[73] Xin Yi, Shuning Zhang, Ziqi Pan, Louisa Shi, Fengyan Han, Yan Kong, Hewu Li, and Yuanchun Shi. 2023. Squeez'In: Private Authentication on Smartphones based on Squeezing Gestures. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. DOI: 10.1145/3544548.3581419

[74] Eunhye Youn, Taejun Kim, and Geehyuk Lee. 2024. WristMenu with Tactons: An Eyes- and Ears-Free Menu with Tactons Describing Menu Items in the Wrist Rotation Space. *International Journal of Human–Computer Interaction* 40, 9. DOI: 10.1080/10447318.2022.2159780