Would You Share the Code? Behavior and Call-Aligned Cardiac Responses in an Unexpected Vishing Attack

Felix Dietz
felix.dietz@ifi.lmu.de
LMU Munich / University of the Bundeswehr Munich
Munich, Germany

Katharina Barlage
katharina.barlage@ifi.lmu.de
LMU Munich
Munich, Germany

Florian Alt
florian.alt@ifi.lmu.de
LMU Munich / University of the Bundeswehr Munich
Munich, Germany

Abstract

Voice phishing (vishing) unfolds across mobile interfaces, where a transaction notification, an incoming call, and the caller's explanation can make each other appear credible. We report a deceptive study of 17 adults examining responses to vishing. Participants received an unexpected transaction-code SMS allegedly from PayPal, followed one minute later by a confederate impersonating PayPal support. Seven were not reached, four ended the call before a code request, four refused, and two disclosed the full code. We collected post-debriefing questionnaires and Polar H10 RR intervals aligned to the call. Among the six participants who reached a disclose-or-refuse decision, median heart rate increased by 2.8 bpm (IQR 1.4–5.0); five showed increases. Changes in ln(RMSSD) were mixed and sensitive to artifact thresholds. We contribute a stage-based account of vishing, clarify the limits of wearable sensing during live security interactions, and derive implications for trusted verification paths, transaction-bound confirmations, and physiology-aware study design.

CCS Concepts

• **Security and privacy** → **Usability in security and privacy**; • **Human-centered computing** → *Empirical studies in HCI*.

Keywords

vishing, voice phishing, social engineering, mobile security, usable security, deception study, heart rate variability, wearable sensing, one-time codes

ACM Reference Format:
Felix Dietz, Katharina Barlage, and Florian Alt. 2026. Would You Share the Code? Behavior and Call-Aligned Cardiac Responses in an Unexpected Vishing Attack. In *25th International Conference on Mobile and Ubiquitous Multimedia (MUM '26), December 01–04, 2026, Glasgow, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3861188.3861228

1 Introduction

Mobile phones bring several trust cues together on one device. A transaction notification can arrive moments before a call, the caller can quote its amount, and the recipient has to decide in real time whether the two events confirm each other or form a coordinated attack. Voice phishing, or *vishing*, exploits exactly this situation. An email can be put aside and read again later; a live call cannot, because pauses, reassurance, and repeated requests reshape the decision while it is being made.

People do answer telephone scams and sometimes hand over sensitive information at scale [31], and generative systems can now run such attacks without a human caller [7]. A single disclosure rate, however, hides very different paths: a person may never receive the call, hang up before the request, refuse explicitly, or disclose after hesitating. Each corresponds to a different point at which a mobile system could intervene.

Observable behavior also tells us little about how the call felt. Wearable sensing offers a way to follow the time course of such an interaction without interrupting it for self-report, but heart rate (HR) and heart rate variability (HRV) also respond to posture, respiration, speech, movement, age, and health, which limits how far a call-aligned change can be read.

We study these questions in a bounded deception study with 17 adults. Under a cover story unrelated to security, participants were sent an SMS containing a six-digit transaction authentication number (TAN), followed roughly one minute later by a confederate call asking for that code. We combine observer-coded behavior, post-debriefing appraisals, and Polar H10 RR intervals. We ask:

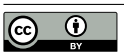
- **RQ1:** How do participants move through an unexpected vishing attempt that begins with an SMS, from reachability to early termination, explicit refusal, or full TAN disclosure?
- **RQ2:** After debriefing, how do participants rate the severity of and their vulnerability to mobile financial fraud, their own capacity to respond, and the ethical acceptability of the study?
- **RQ3:** Which call-aligned HR and HRV patterns can we observe, and which measurement constraints limit their interpretation?

We contribute a stage-based account of a live vishing attempt that does not treat non-answering as evidence of secure judgment; a call-aligned analysis of RR intervals showing a modest, heterogeneous HR increase at the decision stage and no consistent HRV direction; and implications for trusted exit paths, non-transferable authentication, and the instrumentation physiology-aware mobile research needs.

2 Related Work

2.1 Vishing as an Interactive Mobile Attack

Tu et al. ran ten telephone-phishing experiments with 3,000 university participants, showed that recipients answer ethically deployed scam calls, and found that caller-ID presentation affected



This work is licensed under a Creative Commons Attribution 4.0 International License.
MUM '26, Glasgow, United Kingdom
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2791-7/2026/12
https://doi.org/10.1145/3861188.3861228

success [31]. Rahman et al. report the equivalent result for the SMS channel, where a sizeable share of recipients acted on unsolicited text messages [23]. Lain et al. make a related point at organizational scale, showing how much the reported effect of a phishing intervention depends on which denominator and time window are used [18]. Our study is far smaller, but adds a live confederate who could react to resistance, a corroborating SMS, and participant-level temporal sensing.

Vishing is conversational persuasion rather than a static stimulus. Jones et al. analyzed how social engineers apply established persuasion principles during vishing calls [15], and Triantafyllopoulos et al. survey vishing detection research, arguing for human-centered, multimodal approaches [30]. Work on email phishing is instructive too: Butavicius et al. found that time pressure significantly lowered detection accuracy, while the deception cues users are usually told to watch for did not improve discrimination [5]. In a call, the attacker is present and can reinterpret a pause or a refusal immediately, so we keep hanging up, explicit refusal, and disclosure behaviorally distinct.

2.2 Physiological Measurement in Security Interactions

Physiological measures can reveal processes that behavioral accuracy alone cannot. Neupane et al. combined EEG and eye tracking in phishing-detection and malware-warning tasks, illustrating both the promise of multimodal measurement and how much interpretation depends on the task [21]. A chest strap is less obtrusive and can follow a natural phone interaction, but HR and HRV are far less specific: many physical and conversational processes produce the same pattern.

Established guidance recommends five-minute short-term recordings when comparability and physiological interpretation are the goal [17, 19], and RMSSD is the usual choice for short-term variability, with published norms for resting recordings [26]. Ultra-short recordings can correlate with longer resting measurements, especially for RMSSD, but their placement and duration need explicit justification [20]. Those conditions do not make a 30-second speaking interval equivalent to a resting clinical measurement [17]. The Polar H10 agrees closely with ECG-derived RR and HR measures and retains a high detection rate during movement [10, 25], though device validity does not remove artifacts from a particular deployment. We therefore use the H10 as a source of timestamped RR intervals, apply transparent quality rules, and avoid treating HRV as a psychological label.

2.3 Threat Appraisal, Coping, and Deception Ethics

Protection Motivation Theory (PMT) separates threat appraisal, covering perceived severity and vulnerability, from coping appraisal, covering self-efficacy, response efficacy, and response cost; the revised theory also accounts for rewards that sustain maladaptive behavior [24]. PMT is widely used in information-security research, although reported relationships vary across behaviors and designs [12, 27], and Bax et al. show that rewards and response costs are relevant to maladaptive phishing behavior [4]. We use these constructs to describe post-study perceptions, not to explain disclosure: the survey followed both the attack and the debriefing, and only two participants disclosed.

Telling participants in advance that an attack will happen invalidates the behavior under study. Deceptive phishing experiments can improve ecological validity, but require proportional risk, bounded exposure, debriefing, and a real opportunity to withdraw [8, 31]. Distler et al. survey how usable security studies represent risk to participants and treat deception as one option among several [6]. We followed this logic and asked participants directly about ethical concern and whether they would recommend taking part.

The TAN is also an interface object. Current NIST guidance states that out-of-band and one-time-password outputs transferred by hand are not phishing-resistant, because an impostor verifier can relay them to the legitimate service [29]. Kondracki et al. show this relay is already automated on the web, where man-in-the-middle toolkits proxy credentials and second-factor codes to the real site in real time [16]. A caller who asks a person to read a code aloud performs the same relay by hand.

3 Method

3.1 Design and Participants

We ran a single-condition deceptive laboratory study, with participants invited under the cover story that we were testing a new user interface. The invitation mentioned neither the security purpose nor the incoming call. We recruited 17 participants and analyze all of them; nobody withdrew, requested deletion, or was excluded. Table 1 summarizes the sample. Ages ranged from 22 to 61 years, ten participants gave their occupation as student, 15 reported having a PayPal account, and three reported only basic German proficiency. Questionnaires were administered in German; English translations are in Appendices C and D.

Data collection ran in two blocks about six months apart, with five participants (P11–P15) in the first and twelve (P21–P32) in the second, using identical equipment and the same protocol. The blocks differed in season: the first ran in winter at ordinary room temperature, the second during a summer heat period with outdoor temperatures above 30 °C and a warm, non-air-conditioned laboratory. We did not log room temperature and treat this as an uncontrolled between-block difference. Ambient heat raises heart rate and reduces vagally mediated HRV [1], and reporting thermal conditions is standard advice for psychophysiological work [17]; we return to what this does and does not affect in Sections 4.1 and 6.

3.2 Procedure

On arrival, we fitted participants with a Polar H10 chest strap and told them that the previous participant was running late, so they could read or use their phone while they waited. This let the SMS and the call arrive outside any overt task. After a short interval, the telephone number given for the session received the following SMS:

> *Please confirm with the Code 194412 the purchase amount of EUR 49,98 at PAYPAL*

The message was sent in this form, including its unidiomatic English and the German decimal comma. About one minute later,

Table 1: Participant characteristics ($N = 17$). No participants were excluded.

Characteristic	Value
Age, median [IQR], range	28 [24, 30], 22–61
Women / men	12 / 5
Students / other occupations	10 / 7
University degree / school-leaving qualification	9 / 8
PayPal account, yes / no	15 / 2
German, native / basic / other proficiency	11 / 3 / 3
Pre-existing cardiovascular condition, yes	1
Alcohol or other drugs in prior 48 hours, yes	1

a research confederate called and impersonated a PayPal representative, using an ordinary research telephone number. We did not spoof the caller ID, so participants saw an unknown mobile number rather than one resembling PayPal. The confederate said that an unexpected transaction from abroad had been put on hold, repeated the amount, and asked whether the participant had initiated it. When the participant denied it, the confederate said the transaction could be stopped by providing the TAN; when the participant declined, that the TAN was needed to authenticate them. The request was repeated at most twice, and participants were free to end the call at any point.

We monitored the session live against two stop criteria. The first was a heart-rate threshold fixed in the ethics approval, defined per age band: 130 bpm for ages 18–30, 120 bpm for 31–50, and 110 bpm for 51–70, all well below the age-predicted maximum for each band [28]. The second was observational: both researchers followed the call and either could end the session immediately if a participant appeared to be in serious emotional distress. The criterion was deliberately loose, because we preferred stopping a session unnecessarily over continuing one we should have ended. Neither criterion was met.

Immediately afterwards both researchers entered the room, disclosed the study purpose and the deception, explained safer responses to comparable incidents, and pointed out security-preserving behavior where it applied. Participants then received the true study and consent information and could refuse the use of their data or request deletion without losing the agreed compensation. The protocol received institutional ethics approval before recruitment; we omit approval details during double-blind review.

3.3 Measures

Behavior. Each participant had a recorded call duration and an observer note. We applied four rule-based, mutually exclusive categories: *unreached*, where no two-way interaction took place, including an unanswered or muted phone and an unusable number; *early termination*, where the participant answered but ended the call before any TAN disclosure or refusal; *explicit refusal*, where the participant engaged with the request but did not provide the TAN; and *full disclosure*, where the participant said all six digits, continuously or in stages. The scheme deliberately does not treat non-answering as evidence that a participant detected the deception.

Questionnaires. The demographic form recorded age, gender, occupation, education, language background, cardiovascular conditions, recent alcohol or drug consumption, and four statements from the Affinity for Technology Interaction (ATI) scale on its six-point format [9]. The post-study form contained two seven-point items for each of seven PMT-related constructs, asked about PayPal account ownership, and included three seven-point items on ethical concern and on recommending participation to friends or older relatives.

Cardiac data. The sensor logger exported phone-timestamped RR intervals, and Bluetooth disconnects split some sessions across several files, which we reconstructed per participant. Call time marks the recorded start of the call attempt rather than the exact moment of connection. We aligned each trace to the recorded call time ($t = 0$) and kept gaps as gaps. Zero-second attempts appear as a call-time marker only; for an unreached participant with a non-zero duration, the shaded region marks the recorded attempt or ringing interval, not a conversation.

3.4 Analysis

We discarded implausible and artifact-flagged RR intervals rather than interpolating them, so that Bluetooth gaps and rejected beats stay visible as gaps throughout. Participant trajectories show median HR in 10-second bins and ln(RMSSD) from trailing 60-second windows, each shown only where enough valid data supports it. These rolling values expose temporal structure, but their overlapping windows are not independent observations.

The calls that reached a TAN decision lasted 70 to 210 seconds. To compare those six participants on equal terms, we computed median HR and ln(RMSSD) in the 30 seconds before the recorded call time and in the first 30 seconds after it; all six met our coverage criteria. A 30-second RMSSD is an exploratory ultra-short measure: prior validation supports short RMSSD estimates under controlled conditions [20], but a speaking interval is neither stationary nor a resting assessment. We therefore report per-participant changes, medians, IQRs, and direction only - no significance testing - and check whether the direction holds under stricter and looser artifact thresholds. Each 30-second window required at least 80% retained coverage and at least 20 valid adjacent differences; a difference counted only if both intervals were retained, consecutive, and from the same file (to exclude Bluetooth-gap artifacts).

Given the small sample, the behavioral and questionnaire analyses are descriptive. Each PMT construct is the participant-level mean of its two items, reported as a median and IQR across participants; these are descriptive indices, not validated scale scores. To expose how much the denominator matters, we calculate full-disclosure proportions for all participants, for those who answered, and for those with a recorded disclose-or-refuse decision, with Wilson 95 % intervals. We did not run subgroup tests, outcome prediction, or a disclosure-versus-refusal physiology test.

Table 2: Behavioral outcome stages ($N = 17$).

Outcome	*n*	%
Unreached, no two-way interaction	7	41.2
Answered, no TAN decision recorded	4	23.5
Refused TAN explicitly	4	23.5
Disclosed all six digits	2	11.8
Total	17	100.0

4 Results

4.1 Data Quality

All 17 participants had RR data covering the recorded call time. The median flagged-interval rate was 0.4 % (IQR 0.1–1.0; range 0.0–25.4), with P14 the only participant above 5 %, and median retained coverage across each session was 99.2 % (IQR 91.5–99.7; range 57.1–100.0). The lower values reflect Bluetooth gaps rather than removed participants: P31 lost most of the later call interval, though the first decision-stage window still passed our criteria. Appendices A and B show all traces with their missing data rather than smoothing over it.

The recordings also let us check the stop criterion after the fact. Inside the call intervals no participant reached the threshold for their age band; the highest values are 128.5 bpm for P15 and 127.9 bpm for P30, both against 130 bpm. Four participants briefly exceeded their threshold outside the call. The longest episode is P30, whose HR stayed between 135 and 152 bpm for about 80 seconds starting roughly 20 seconds after her call ended, coinciding with the researchers entering to debrief her.

We also inspected pre-call levels for an effect of the different ambient conditions. Median HR from five minutes to 30 seconds before the call was 89.0 bpm (IQR 81.9–95.2) in the winter block and 87.6 bpm (IQR 83.3–95.8) in the summer block. We report this for transparency only: with five and twelve participants, and season confounded with everything else that differed between blocks, the comparison is uninformative.

4.2 Progress Through the Call

Table 2 shows the attack funnel. Seven of the 17 participants were unreached: five did not answer, one had the phone muted and noticed neither the message nor the call, and one had given an unusable number. Ten answered. Four ended the call before a TAN decision was recorded: two hung up after hearing "PayPal", one ended the call immediately after greeting, and one said they had no time. Four refused the request explicitly, and two disclosed all six digits.

The answer rate was 10/17 (58.8 %, Wilson 95 % CI [36.0 %, 78.4 %]). Full disclosure was 2/17 across all participants (11.8 %, [3.3 %, 34.3 %]), 2/10 among those who answered (20.0 %, [5.7 %, 51.0 %]), and 2/6 among those with a recorded disclose-or-refuse decision (33.3 %, [9.7 %, 70.0 %]). These intervals are wide and indicate imprecision rather than population prevalence. Among participants who answered, the median call duration was 73.5 seconds (IQR 21.0–100.5; range 3–210).

Table 3: Post-study construct indices. Each index is the participant-level mean of two items on a 1–7 scale. Higher values indicate more of the named construct.

Construct	Median [IQR]
Severity	6.5 [6.5, 7.0]
Vulnerability	3.5 [3.0, 4.5]
Maladaptive intrinsic rewards	3.5 [2.5, 4.5]
Maladaptive extrinsic rewards	5.0 [5.0, 5.5]
Self-efficacy	4.0 [4.0, 5.5]
Response efficacy	6.0 [5.0, 6.5]
Response cost	4.0 [3.5, 5.0]

The observer notes reveal behavior that a binary outcome hides. One discloser first gave four digits, paused for confirmation, then said the final two quietly, and only after all six had been transferred did she ask for an email confirmation: hesitation and an attempt to verify, but not an incomplete disclosure. The other discloser reported not having a PayPal account. Neither case supports a subgroup claim; together they show that neither hesitation nor an account mismatch necessarily stopped the scripted interaction. Among the refusals, P30 did not simply decline but said she would call back in five minutes, a verification attempt we return to in Section 5.3.

4.3 Post-Study Appraisals and Acceptability

Table 3 reports the descriptive indices. Participants rated the consequences of criminals accessing their bank data through the phone as highly severe (median 6.5/7) and protective responses as effective (6.0/7). Perceived vulnerability was lower (3.5/7) and self-efficacy sat at the scale midpoint (4.0/7). That gap is descriptively notable: participants endorsed protective actions much more uniformly than their own ability to recognize threats or configure protections. Because the survey came after both the attack and the debriefing, it cannot be read as a pre-existing cause of behavior. The four ATI items, with the two reverse-keyed items recoded, gave a median of 3.25 on the six-point scale (IQR 2.50–3.75; range 1.75–6.00), slightly below the midpoint; we report this as sample description only.

Ethical concern was not absent. The item "I found the conduct of this study ethically questionable" had a median of 3 (IQR 2–4): ten participants selected 1–3, three the midpoint, and four selected 5–7. Fifteen participants selected 5–7 for recommending the study to friends (median 6, IQR 5–7), and eleven for recommending it to grandparents or other older relatives (median 6, IQR 3–7). Recommendation and ethical concern are not two ends of one scale; a participant could see educational value and still object to the deception.

4.4 Call-Aligned Heart Rate and HRV

Figure 1 shows HR and ln(RMSSD) for the six participants who reached a TAN decision, from four minutes before to four minutes after the recorded call time. For a standardized comparison, Table 4 gives equal 30-second windows before and after that time. Median HR increased for five of the six participants and decreased for one, ranging from −1.6 to +6.6 bpm with a median of +2.8 bpm (IQR

1.4–5.0). The direction was robust to the artifact threshold: the median change ranged from +2.3 to +2.8 bpm across the 15 %, 20 %, and 25 % settings.

Ln(RMSSD) showed no common direction: three participants increased and three decreased, with changes from −0.78 to +0.45 and a median of 0.06 (IQR −0.22 to 0.27). The median also moved from −0.16 under a 15 % rule to +0.12 under a 25 % rule. In summary, the direction is mixed and sensitive to the artifact threshold; HRV did not increase or decrease as a group.

The trajectories in Figure 1 add context. Participants entered the call with very different HR and RMSSD levels, and part of that spread is environmental rather than psychological, since the two blocks were recorded in very different thermal conditions. Some calls coincide with transient changes, but their shapes differ: P23's HR rose during the call interval while P32's stayed comparatively stable, and P30 showed a later and much larger excursion after her call had already ended. The jump in the refusal median about 1.5 minutes after the call reflects P15 and P30 rather than the group, as P31 and P32 stayed flat. The rolling HRV curves also move before and after $t = 0$, which illustrates why picking a peak or treating every point as independent would mislead. These traces answer RQ3 descriptively: call-aligned dynamics are observable at the individual level, but they vary in shape and timing rather than following a common pattern.

5 Discussion

5.1 Security Is a Funnel

The largest outcome group was neither disclosure nor refusal but non-exposure: seven participants never entered a two-way interaction. Treating all non-disclosers as having recognized the scam would overstate how much secure judgment we observed, while reporting only the six with a recorded TAN decision would ignore a behavior that matters in everyday mobile use, namely ending an unexpected call before the sensitive request develops. Our three denominators answer three questions: deployment yield, risk after engagement, and what happened once a TAN decision was documented. Prior work often makes this choice implicitly. Harris et al. excluded the 20.5 % of enrolled participants who never answered after five attempts [13], so their 31.18 % success rate describes risk after engagement.

This maps directly onto intervention design. Caller authentication and screening act before the call is answered; conversational warnings and trusted callback paths act after answering but before the request; transaction-confirmation design acts at authorization. A single "vishing success rate" collapses all three and can make unlike systems look comparable.

Our caller ID matters here. The confederate called from an ordinary mobile number with no spoofing, the weakest possible legitimacy cue at the screening stage. Tu et al. found that a spoofed caller ID significantly increased success [31], so our answer rate and disclosure counts should be read as a lower bound for an attacker who also controls the displayed number.

5.2 What the Cardiac Data Do and Do Not Show

At the decision stage we see a modest HR acceleration for most participants around the call, while ln(RMSSD) changes split evenly and shift with the artifact threshold. HR and HRV are related but not interchangeable, and neither measure tells us whether a participant was frightened, suspicious, cognitively engaged, speaking more, or simply shifting posture. P30 illustrates the problem: her largest HR excursion did not occur during the call at all, but in the minute after it, when the researchers entered to debrief her.

Synchronization determined what we could claim. Missing speech markers and Bluetooth fragments limited how far we could interpret individual events, so future mobile-security studies should log notification delivery, ring, answer, request, disclosure or refusal, hang-up, and debrief on a shared clock. Audio turn boundaries and respiration or movement measures would help separate conversational mechanics from cardiac change, and analysis should stay at the participant level to avoid pseudoreplication from dense sensor samples.

None of this justifies an interface that labels a user as "stressed" from HR or HRV and blocks a transaction, which would face uncertain validity, sensitive health data, false positives, and autonomy concerns. A more defensible direction is consented, on-device use of physiology as one uncertain contextual signal alongside interaction events, with fallbacks that do not depend on a physiological classification. In our data the strongest actionable information is still behavioral: an unexpected request for a transferable code, arriving alongside a notification that appears to corroborate it.

5.3 From Readable Codes to Trusted Exits

The staged disclosure corrects a tempting but unsafe reading of partial speech. The participant paused, waited for a response, and lowered her voice, and she still transferred the complete credential. That is hesitation, not residual cryptographic protection: in a live call the attacker can wait, reassure, or repeat, and security design should not rely on a user stopping halfway through a short code. More fundamentally, a TAN that can be read aloud is a transferable bearer secret; the web equivalent of our confederate is already tooled and automated [16], and nothing about a spoken relay is harder. Mobile financial services should prefer cryptographic confirmation that binds the user, the application, the transaction, and the destination.

Eight of the ten participants who answered did not disclose. We did not interview them, so we cannot attribute their outcomes to any particular cue, but the behavior shows a design opportunity: make leaving the caller's narrative easy and actionable. P30 makes this concrete. Rather than refusing outright, she told the caller she would call back in five minutes. That is precisely the move a trusted exit should support, and today it is left entirely to the user, who would have to find a legitimate number herself without help from the app that generated the notification. After a sensitive financial notification, a service could instead offer an authenticated "review activity" action, a verified support entry point, or a one-tap account lock that does not depend on contact information supplied by the caller.

Such interventions need care: cross-app inference about calls and financial notifications has privacy implications, and users habituate to repeated warnings quickly enough that the effect is measurable after a handful of exposures [2]. Research can test which component should mediate these signals, when friction is warranted, and how

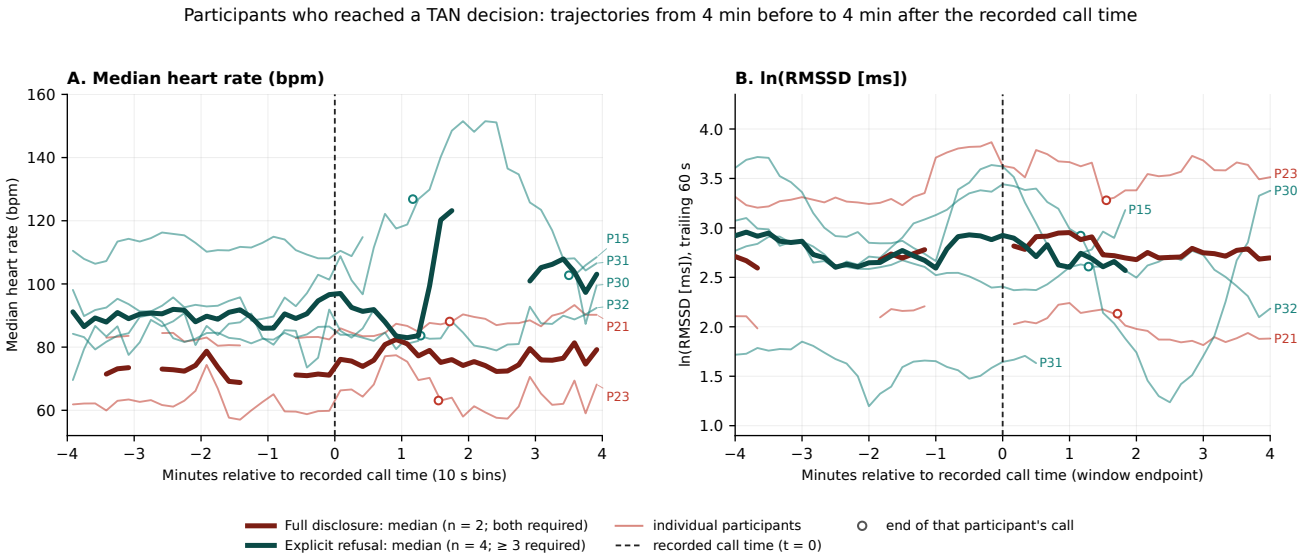


Figure 1: Heart rate (A) and ln(RMSSD) (B) for the six participants who reached a TAN decision, from four minutes before to four minutes after the recorded call time (dashed line). Thin lines are individual participants and open circles mark the end of each call. Thick lines are outcome medians (red: full disclosure, $n = 2$; teal: explicit refusal, $n = 4$), drawn only where both disclosers or at least three refusers have data. Gaps are not interpolated. The colors aid traceability; we do not infer an outcome-group comparison. The 30-second window values are given in Table 4.

to process the context locally. The survey adds a complementary signal: response efficacy was high while self-efficacy sat near the midpoint, so advice may be more useful when it supplies a concrete next action than when it only describes danger. "Hang up and open the app to review this transaction" is executable; "be careful of fraud" leaves the user inside the attacker's frame. Training alone is unlikely to close this gap. In Hashmi et al.'s live-call test, participants trained on the persuasion tactics of vishing disclosed about as often as those given general awareness training, though the authors present this as a preliminary result from a small sample [14].

5.4 Deception and Participant Acceptability

Most participants would recommend the study to friends and fewer to older relatives, while four agreed that its conduct was ethically questionable, one at the strongest response option. These results do not certify our procedure as ethical; they show that reactions after a bounded deception, immediate debriefing, security education, and a withdrawal opportunity were heterogeneous. The gap between the two recommendation items is worth noting given that older adults are among the groups most affected by targeted fraud [22]: participants were most hesitant about exposing exactly the population such a study might inform.

Future deceptive security studies should report this dissent instead of collapsing acceptability into a mean. The protections we used are part of the method: a bounded number of requests, continuous safety monitoring against a predefined heart-rate criterion and a researcher's discretion to stop the call, immediate disclosure, compensation independent of behavior, and the option to delete the data. Physiological monitoring adds a second layer of sensitivity, because participants must be told after the deception not only which behavior was studied but also which biosignal was collected and how it will be used.

6 Limitations and Future Work

Our sample is small, with 17 participants, six TAN decisions, and two disclosures, and cannot support prevalence estimates, demographic comparisons, or disclosure prediction. It includes 12 women and 10 students and is not representative of mobile-banking users. We also collected data in two blocks six months apart, in winter and during a summer heat period. Ambient heat raises heart rate and reduces vagally mediated HRV [1], so the blocks are not thermally comparable and absolute levels should not be compared across them. This affects our decision-stage analysis far less, because both windows come from the same participant in the same room 30 seconds apart, so any thermal offset cancels in the difference; but five of the six decision-stage participants come from the summer block, so we cannot rule out that the size of a call-related change differs under thermal load.

The procedure uses one brand, one amount, one code, one caller, and one fraud story in a laboratory waiting situation, and because we did not spoof the caller ID our results speak to an attacker without that capability. Seven participants were never exposed to spoken manipulation; this deployment friction is meaningful in itself but reduces the amount of conversational data.

The physiological analysis is exploratory. We have no standardized resting baseline and no legitimate-call control, and movement, posture, and anticipation after the SMS are uncontrolled. One participant reported a pre-existing cardiovascular condition and one reported alcohol or drug use within 48 hours. A 30-second RMSSD

is shorter than the conventional five-minute standard and we use it only for equal-window description.

Most importantly, we cannot separate a response to the fraud from a response to being called at all. An unexpected incoming call is a salient event in its own right, and salient stimuli reliably produce a cardiac response regardless of what they turn out to mean [11]. Ringing, deciding whether to answer, picking up, and beginning to speak all fall inside the window we analyze, so the small HR increase is at least as consistent with the arousal of answering an unknown number as with anything specific to the fraud. A participant who is simply startled by their phone and one who senses that something is wrong produce the same measurement here. Talking on the phone changes cardiac measures as well. In healthy volunteers, a 20-minute call raised SDNN and high-frequency power compared with the periods before and after it, an effect the authors could not separate from speaking itself [3]. A call can therefore shift HRV in the opposite direction to what stress alone would predict.

That is the strongest reason for a benign-call control, without which a call-aligned change cannot be attributed to the fraud rather than to the call. A follow-up study should run that comparison directly, ideally within participants with condition order counterbalanced, treating the benign call as the reference rather than a resting baseline; suspicious-call and no-call conditions would further separate general wariness of unknown numbers from a response to the specific request. Respiration and movement sensing with post-event interviews would enable stronger cross-validation. System studies should then test trusted callback paths, transaction-bound confirmations, and privacy-preserving context-aware interventions, rather than assuming that stronger warnings or physiological classification alone will stop a persuasive caller.

7 Conclusion

An unexpected vishing call is a sequence of opportunities to disengage, not a single disclosure decision. In our 17-person deception study, seven participants were never reached, four ended the call before a TAN decision was recorded, four refused explicitly, and two disclosed the entire code, one of them hesitantly and in stages, which shows why hesitation should not be mistaken for protection. Call-aligned wearable data added a second perspective: five of six decision-stage participants showed a small early HR increase, while ln(RMSSD) changes split evenly and shifted with our analysis choices. For mobile security design, the results highlight the value of moving verification out of the caller-controlled interaction and into a trusted service channel, and of transaction approval that identifies the recipient directly rather than depending on a code that can be spoken aloud.

Acknowledgments

The presented work received funding from the Technology Research Center of the Bundeswehr (Voice of Wisdom). dtec.bw is funded by the European Union - NextGenerationEU. The authors used AI-Tools such as Grammarly, ChatGPT, and Claude to improve grammar, punctuation, and vocabulary.

References

[1] Oriol Abellán-Aynés, Pedro Manonelles, and Fernando Alacid. 2021. Cardiac Parasympathetic Withdrawal and Sympathetic Activity: Effect of Heat Exposure on Heart Rate Variability. *International Journal of Environmental Research and Public Health* 18, 11 (2021). doi:10.3390/ijerph18115934

[2] Bonnie Brinton Anderson, C. Brock Kirwan, Jeffrey L. Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2883–2892. doi:10.1145/2702123.2702322

[3] Ryszard Andrzejak, Rafal Poreba, Malgorzata Poreba, Arkadiusz Derkacz, Robert Skalik, Pawel Gac, Boguslaw Beck, Aleksandra Steinmetz-Beck, and Witold Pilecki. 2008. The influence of the call with a mobile phone on heart rate variability parameters in healthy volunteers. *Industrial health* 46, 4 (2008), 409–417. doi:10.2486/indhealth.46.409

[4] Samantha Bax, Tanya McGill, and Val Hobbs. 2021. Maladaptive behaviour in response to email phishing threats: The roles of rewards and response costs. *Computers & Security* 106 (2021), 102278. doi:10.1016/j.cose.2021.102278

[5] Marcus Butavicius, Ronnie Taib, and Simon J. Han. 2022. Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. *Computers & Security* 123 (2022), 102937. doi:10.1016/j.cose.2022.102937

[6] Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. 2021. A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research. *ACM Trans. Comput.-Hum. Interact.* 28, 6, Article 43 (Dec. 2021), 50 pages. doi:10.1145/3469845

[7] João Figueiredo, Afonso Carvalho, Daniel Castro, Daniel Gonçalves, and Nuno Santos. 2025. On the Feasibility of Fully AI-automated Vishing Attacks. arXiv:2409.13793 [cs.CR] https://arxiv.org/abs/2409.13793

[8] Peter Finn and Markus Jakobsson. 2007. Designing ethical phishing experiments. *IEEE Technology and Society Magazine* 26, 1 (2007), 46–58. doi:10.1109/MTAS.2007.335565

[9] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. arXiv:https://doi.org/10.1080/10447318.2018.1456150 doi:10.1080/10447318.2018.1456150

[10] Rahel Gilgen-Ammann, Theresa Schweizer, and Thomas Wyss. 2019. RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. *European journal of applied physiology* 119, 7 (2019), 1525–1532.

[11] Frances K Graham and Rachel K Clifton. 1966. Heart-rate change as a component of the orienting response. *Psychological bulletin* 65, 5 (1966), 305.

[12] Steffi Haag, Mikko Siponen, and Fufan Liu. 2021. Protection Motivation Theory in Information Systems Security Research: A Review of the Past and a Road Map for the Future. *SIGMIS Database* 52, 2 (April 2021), 25–67. doi:10.1145/3462766.3462770

[13] Ian G. Harris, Ali Derakhshan, and Marcel Carlsson. 2020. A Study of Targeted Telephone Scams Involving Live Attackers. In *Socio-Technical Aspects in Security and Trust: 10th International Workshop, STAST 2020, Virtual Event, September 14, 2020, Revised Selected Papers*. Springer-Verlag, Berlin, Heidelberg, 63–82. doi:10.1007/978-3-030-79318-0_4

[14] Sumair Ijaz Hashmi, Niklas George, Eimaan Saqib, Fatima Ali, Nawaal Siddique, Shafay Kashif, Shahzaib Ali, Nida Ul Habib Bajwa, and Mobin Javed. 2023. Training Users to Recognize Persuasion Techniques in Vishing Calls. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 294, 8 pages. doi:10.1145/3544549.3585823

[15] Keith S. Jones, Miriam E. Armstrong, McKenna K. Tornblad, and Akbar Siami Namin. 2020. How social engineers use persuasion principles during vishing attacks. *Information and Computer Security* 29, 2 (12 2020), 314–331. arXiv:https://www.emerald.com/ics/article-pdf/29/2/314/785860/ics-07-2020-0113.pdf doi:10.1108/ICS-07-2020-0113

[16] Brian Kondracki, Babak Amin Azad, Oleksii Starov, and Nick Nikiforakis. 2021. Catching Transparent Phish: Analyzing and Detecting MITM Phishing Toolkits. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (Virtual Event, Republic of Korea) *(CCS '21)*. Association for Computing Machinery, New York, NY, USA, 36–50. doi:10.1145/3460120.3484765

[17] Sylvain Laborde, Emma Mosley, and Julian F. Thayer. 2017. Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research – Recommendations for Experiment Planning, Data Analysis, and Data Reporting. *Frontiers in Psychology* Volume 8 - 2017 (2017). doi:10.3389/fpsyg.2017.00213

[18] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. 2022. Phishing in Organizations: Findings from a Large-Scale and Long-Term Study. In *2022 IEEE Symposium on Security and Privacy (SP)*. 842–859. doi:10.1109/SP46214.2022.9833766

[19] Marek Malik, J. Thomas Bigger, A. John Camm, Robert E. Kleiger, Alberto Malliani, Arthur J. Moss, and Peter J. Schwartz. 1996. Heart rate variability: Standards of

measurement, physiological interpretation, and clinical use. *European Heart Journal* 17, 3 (03 1996), 354–381. arXiv:https://academic.oup.com/eurheartj/article-pdf/17/3/354/1312587/17-3-354.pdf doi:10.1093/oxfordjournals.eurheartj.a014868

[20] M. Loretto Munoz, Arie van Roon, Harriëtte Riese, Chris Thio, Emma Oostenbroek, Iris Westrik, Eco J. C. de Geus, Ron Gansevoort, Joop Lefrandt, Ilja M. Nolte, and Harold Snieder. 2015. Validity of (Ultra-)Short Recordings for Heart Rate Variability Measurements. *PLOS ONE* 10, 9 (09 2015), 1–15. doi:10.1371/journal.pone.0138921

[21] Ajaya Neupane, Md. Lutfor Rahman, Nitesh Saxena, and Leanne Hirshfield. 2015. A Multi-Modal Neuro-Physiological Study of Phishing Detection and Malware Warnings. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 479–491. doi:10.1145/2810103.2813660

[22] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. 2017. Dissecting Spear Phishing Emails for Older vs Young Adults: On the Interplay of Weapons of Influence and Life Domains in Predicting Susceptibility to Phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6412–6424. doi:10.1145/3025453.3025831

[23] Md Lutfor Rahman, Daniel Timko, Hamid Wali, and Ajaya Neupane. 2023. Users Really Do Respond To Smishing. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy* (Charlotte, NC, USA) *(CODASPY '23)*. Association for Computing Machinery, New York, NY, USA, 49–60. doi:10.1145/3577923.3583640

[24] Ronald W. Rogers. 1983. Cognitive and Physiological Processes in Fear Appeals and Attitude Change: A Revised Theory of Protection Motivation. In *Social Psychophysiology: A Sourcebook*. Guilford Press, New York, NY, USA, 153–176.

[25] Marcelle Schaffarczyk, Bruce Rogers, Rüdiger Reer, and Thomas Gronwald. 2022. Validity of the Polar H10 Sensor for Heart Rate Variability Analysis during Resting State and Incremental Exercise in Recreational Men and Women. *Sensors* 22, 17 (2022). doi:10.3390/s22176536

[26] Fred Shaffer and J. P. Ginsberg. 2017. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health* Volume 5 - 2017 (2017). doi:10.3389/fpubh.2017.00258

[27] Teodor Sommestad, Henrik Karlzén, and Jonas Hallberg. 2015. A meta-analysis of studies on protection motivation theory and information security behaviour. *International Journal of Information Security and Privacy (IJISP)* 9, 1 (2015), 26–46.

[28] Hirofumi Tanaka, Kevin D Monahan, and Douglas R Seals. 2001. Age-predicted maximal heart rate revisited. *JACC* 37, 1 (2001), 153–156. doi:10.1016/S0735-1097(00)01054-8

[29] David Temoshok, Yee-Yin Choong, Andrew Regenscheid, Ryan Galluzzo, James Fenton, Justin Richer, and Naomi Lefkovitz. 2025. *NIST SP 800-63B-4:Digital Identity Guidelines - Authentication and Authenticator Management.* Technical Report. doi:10.6028/NIST.SP.800-63B-4

[30] Andreas Triantafyllopoulos, Anika A. Spiesberger, Iosif Tsangko, Xin Jing, Verena Distler, Felix Dietz, Florian Alt, and Björn W. Schuller. 2025. Vishing: Detecting social engineering in spoken communication — A first survey & urgent roadmap to address an emerging societal challenge. *Computer Speech & Language* 94 (2025), 101802. doi:10.1016/j.csl.2025.101802

[31] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. 2019. Users really do answer telephone scams. In *Proceedings of the 28th USENIX Conference on Security Symposium* (Santa Clara, CA, USA) *(SEC'19)*. USENIX Association, USA, 1327–1340.

A Participant-Level Call-Aligned Heart Rate

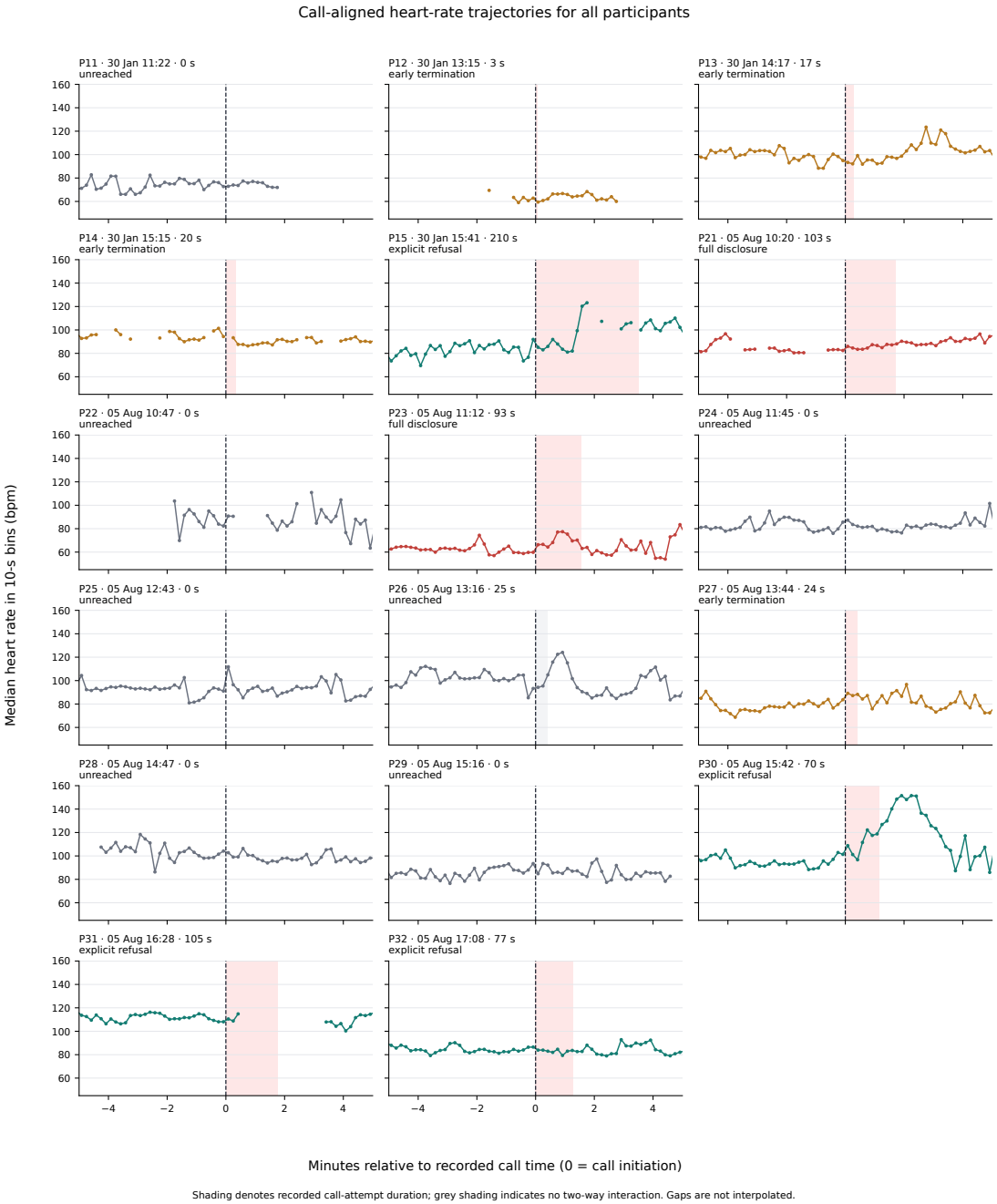


Figure 2: Call-aligned HR trajectories for all 17 participants. Gaps are retained and not interpolated.

B Participant-Level Call-Aligned HRV

Figure 3 shows ln(RMSSD) in trailing 60-second windows. The horizontal position is the window endpoint, so the value at $t = 0$, where present, summarizes the preceding minute. A window is shown only when retained intervals cover at least 80 % of its duration and it contains at least 40 valid adjacent differences; windows that fail these criteria are absent. Line colors follow Appendix A.

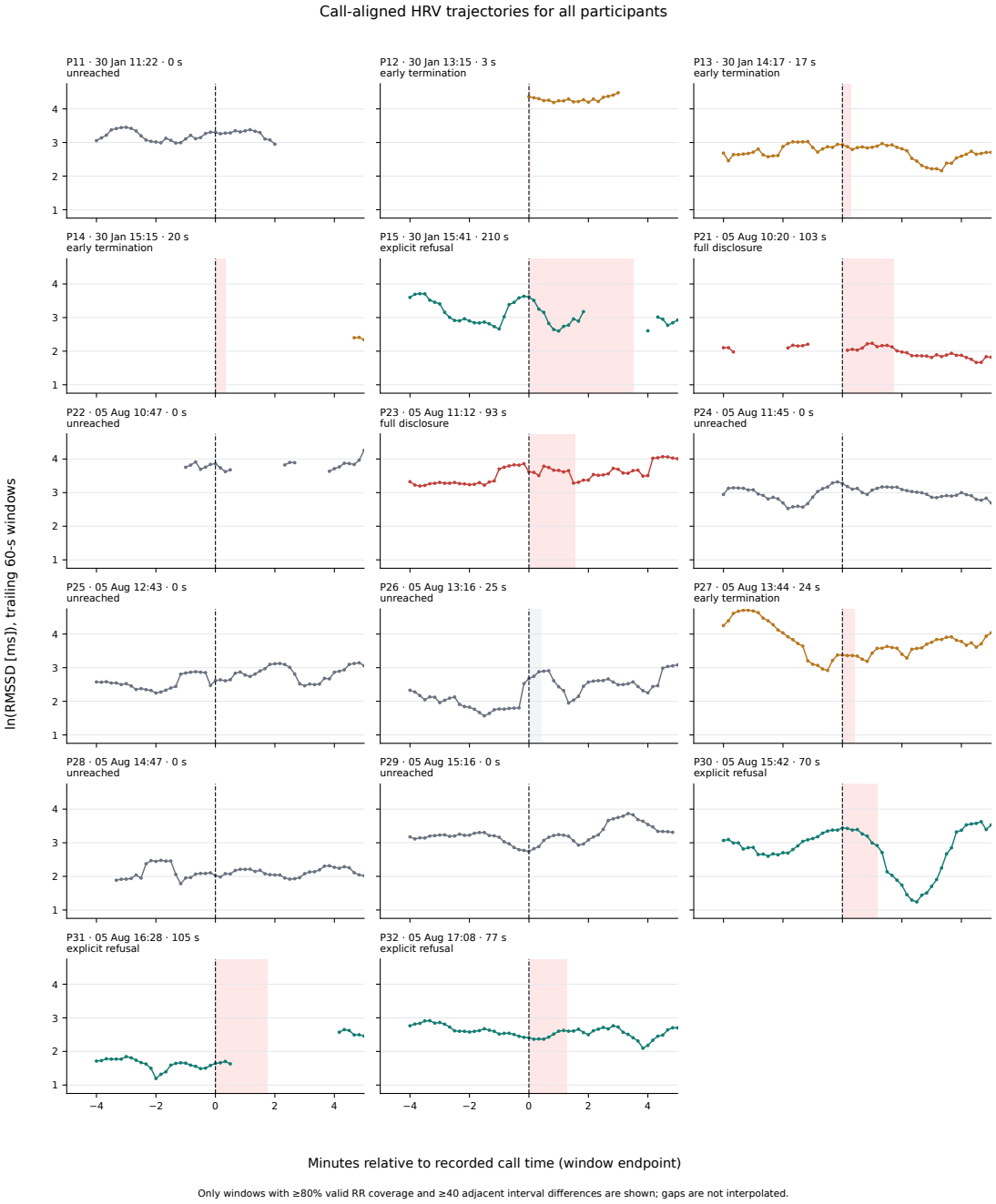


Figure 3: Call-aligned ln(RMSSD) trajectories for all 17 participants. We show only 60-second windows with at least 80 % retained RR coverage and at least 40 valid adjacent differences.

Table 4: Thirty-second standardized windows for the six decision-stage participants. Δ is the first 30 seconds after the recorded call time minus the preceding 30 seconds.

		Median HR (bpm)			
ID	Outcome	Pre	Call	Δ	Δ ln(RMSSD)
P15	Refused	79.1	84.6	5.5	−0.78
P21	Disclosed	83.0	84.2	1.2	0.45
P23	Disclosed	59.2	65.9	6.6	0.28
P30	Refused	99.8	103.4	3.6	−0.11
P31	Refused	108.5	110.4	1.9	−0.26
P32	Refused	85.3	83.7	−1.6	0.24
Median [IQR] of the changes				2.8 [1.4, 5.0]	0.06 [−0.22, 0.27]

C English Translation of the Demographic Questionnaire

This appendix is an English translation of the German instrument we administered at the start of the study. It reproduces item content and response options, not the original page layout.

C.1 Participant Identifier and General Information

- Participant ID: free-text field.
- Age: free-text field.
- Gender: male; female; diverse; prefer not to say.
- Occupation: free-text field.
- Highest educational qualification: free-text field.

C.2 Language Proficiency

- What is your native language? Free-text field.
- How would you rate your German proficiency? Basic knowledge; good knowledge; fluent; business-fluent; native language.

C.3 Physical Condition

- Do you have any pre-existing cardiovascular conditions (e.g., cardiac arrhythmia)? Yes / No.
- Have you consumed alcohol or other drugs within the last 48 hours? Yes / No.

C.4 Technology Interaction

These four items are taken from the Affinity for Technology Interaction (ATI) scale [9]. Instruction: "The following questions concern your interaction with technical systems. With 'technical systems' we mean apps and other software applications as well as complete digital devices (e.g., mobile phone, computer, television, car navigation). Please indicate how strongly you agree with the following statements."

The response scale was: 1 = strongly disagree; 2 = mostly disagree; 3 = somewhat disagree; 4 = somewhat agree; 5 = mostly agree; 6 = fully agree.

(1) I like to engage with technical systems in greater detail.
(2) I like to try out the functions of new technical systems.
(3) It is enough for me that a technical system works; I do not care how or why. (reverse-keyed)
(4) It is enough for me to know the basic functions of a technical system. (reverse-keyed)

D English Translation of the Post-Study Questionnaire

This appendix is an English translation of the German instrument we administered after the debriefing. Unless otherwise stated, items used a seven-point scale: 1 = strongly disagree; 2 = disagree; 3 = somewhat disagree; 4 = neither agree nor disagree; 5 = somewhat agree; 6 = agree; 7 = fully agree.

D.1 Account Ownership

- I have a PayPal account. Yes / No.

D.2 Severity

(1) If criminals gained access to my bank data through my smartphone, this would have serious consequences for me.
(2) Becoming a victim of bank fraud through my smartphone would cause me considerable stress and anxiety.

D.3 Vulnerability

(1) I believe there is a high risk that attackers could gain access to my bank data.
(2) I feel highly susceptible to becoming a victim of telephone banking fraud or identity theft.

D.4 Maladaptive Intrinsic Rewards

(1) Not actively paying attention to the security of my smartphone and banking apps saves me stress and time.
(2) Ignoring security warnings about suspicious calls or text messages makes my everyday life easier.

D.5 Maladaptive Extrinsic Rewards

(1) It seems common for users not to actively protect their smartphones against bank fraud.
(2) It seems common to trust the default settings provided by the smartphone manufacturer or banking apps.

D.6 Self-Efficacy

(1) I am confident that I can recognize and prevent fraudulent calls and text messages.
(2) I know exactly how to change my smartphone settings to protect my bank data from fraud.

D.7 Response Efficacy

(1) Additional security measures on my smartphone (e.g., entering an additional code or using secure PINs) protect me effectively against financial fraud.
(2) Being cautious and verifying the authenticity of bank-related calls substantially reduces my risk of becoming a victim of fraud.

D.8 Response Cost

(1) Implementing and maintaining secure smartphone settings for banking requires considerable time and effort.
(2) Constantly checking calls and TAN requests is burdensome and disrupts my daily routine.

D.9 General Study Evaluation

(1) I found the conduct of this study ethically questionable.
(2) I would recommend participation in this study to my friends.
(3) I would recommend participation in this study to my grandparents (or older relatives).