AI Companions and the Social Shift from Control to Support

Aleksandr Kolomiets
Independent researcher
ORCID 0009-0008-5153-5546

Abstract

We propose a two-level architecture for AI-based companions that aims to reduce interpersonal and collective violence while preserving users' autonomy. At the micro-level, a personal AI companion runs on the user's device, tracks emotional tension in everyday communication and offers "soft" de-escalation prompts. Its core design principle is an autonomy-preserving gate: the system intervenes actively only when it detects a combination of high stress and an explicit intent to harm oneself or others, and only if the user requests help. In all other cases the companion stays in an observational mode and learns from episodes of successful self-regulation.

At the macro-level, we embed this companion into a governance framework consisting of an ethics filter E (platform-level moderation and down-ranking of radical content) and institutional response R (how quickly and in what form public institutions react to spikes of violence). A supranational charter sets corridors for E and R, audits providers and coordinates emergency adjustments of these parameters when key indicators cross a predefined threshold.

We build an agent-based model that links individual-level prevention to societal-level outcomes. The model simulates how different combinations of E, R and companion coverage affect the prevalence of protest and violent behaviour. We calibrate the model using data from V-Dem, Pew and ACLED for 2022–2024 and show that the strongest "dampener" of violence is the quality of democratic institutions, while religious fractionalisation alone does not lead to higher violence. We also provide a first-pass macroeconomic estimate: reallocating up to 0.7 % of global GDP from military spending to AI-companion infrastructure can be net-beneficial under conservative assumptions (SIPRI, World Bank, OECD).

Scanning the (E, R) grid in the agent-based model shows a monotonic reduction in long-run violence as either platform filtering or institutional response increases, with a threshold-like "rollback zone" when both are too weak. The pattern is robust to different initial seeds and

random graph realisations; all code, data and figures are fully reproducible from an open repository (DOI: 10.5281/zenodo.17390730).

Normatively, the companion is designed as a "mirror with support" rather than an instrument of behavioural control. It helps users separate their own needs from imposed expectations and treats bodily signals of violence and gratitude as potentially measurable indicators of the quality of social interaction. We outline operational protocols, validation plans (ABA / RCT, transferability, measurement invariance) and a governance charter compatible with existing human-rights and AI-ethics frameworks. The goal is not to promise a quick "end of wars", but to set up a realistic, auditable trajectory in which supportive systems gradually dominate controlling ones and the default response to tension shifts from coercion to mutual support.

1 Introduction

Digital assistants have already learned to schedule our day, reply to emails and even sustain small talk. A natural next step is an AI companion for violence prevention: a system that unobtrusively tracks emotional tension, suggests ways to reduce aggression towards oneself and others, and thereby helps prevent escalation—from household quarrels to street protests.

The aim of the proposed architecture is to launch a long-term cultural drift in which sensitivity to violence and the preference for non-violent solutions gradually increase, through systematic support of the user's sense of "being right" and having a legitimate point of view. We use a century-long diffusion horizon (by analogy with electrification and mass literacy) as a realistic benchmark rather than a rigid forecast. At the same time, the key capability-gain indicators—reductions in Frustration and Tension, growth of participation and productivity with less external interference—can be piloted already today (see Section 5.1 and Appendix E). In the long run such a drift should also reduce the probability of large-scale violent conflict, but in this paper we focus on near-term, measurable effects.

Mass deployment of such a service creates an ambivalent picture. On the one hand, we expect improved mental well-being and higher productivity [1]. On the other hand, there is a risk of polarisation and violent protest if algorithms unintentionally push users towards the most resonant—and therefore often radical—messages and groups [2, 3].

At the level of individual experience, however, “violence” rarely appears as an abstract moral category. It is more often the end point of a prolonged regulatory failure: a collapse of the person’s ability to reconcile bodily tension, emotional impulses, social expectations and their own sense of what is right. The subject arrives in a split state, where the body is already mobilised for fight or flight, the narrative about self and others has narrowed to a few hostile scripts, and available actions appear either destructive or humiliating. In such states, access to non-violent options is not simply a matter of “knowing the rules”: semantic blind spots and stress-induced tunnel vision make alternative framings literally hard to see. The proposal in this paper is therefore not to “teach better norms” in the abstract, but to design AI-mediated support that lowers this regulatory load, re-opens the field of interpretation and makes non-violent, dignity-preserving responses realistically accessible again.

Anthropological Foundations of Violence and Regulatory Collapse

Violence, in this framework, is not conceptualized as a moral failure but as a form of regulatory collapse.

When tension accumulates beyond the system’s processing capacity, the agent transitions into a survival-oriented regulatory mode. This state is characterised by:

- cognitive narrowing;
- semantic rigidity and reduced interpretive flexibility;
- misalignment between affect, bodily tension and behavioural impulse;
- impaired capacity to hold complex or ambivalent experience.

We refer to this condition as **the split-state**, where regulatory subsystems that normally cooperate (somatic markers, needs recognition, meaning-mapping, intention formation) lose synchrony.

Such fragmentation increases the likelihood of reactive behaviour, misperception, interpersonal misalignment and escalation.

This anthropological perspective is essential for modelling how micro-level tension states translate into interpersonal or societal instability.

To address this, we propose a **two-level system**.

1. **Lower level: personal AI companion.**

The companion runs locally on the user's device, tracks emotional markers, offers gentle techniques for reducing tension and a "funnel" of behavioural prompts. The key mechanism is the **Autonomy-Preserving Gate (AP-Gate)** [4]. The algorithm intervenes actively only when it detects a combination of (i) high tension and (ii) an intention of violent action towards self or others, and when the user explicitly asks for help. In all other situations the companion remains in observer mode and logs episodes of successful self-regulation. A detailed intervention cascade and the logic of AP-Gate are described in Appendix G.2; it is triggered only when both high risk and explicit user consent are present.

2. **Upper level: Ethics Charter and oversight.**

A supranational, cross-sectoral structure (regulators, NGOs, academia, providers) that:

- sets the level of *E* (strictness of the radical-content filter) and the minimally acceptable level of *R* (institutional reaction to spikes of violence) [5];
- carries out regular external audits, ensuring that AP-Gate does not infantilise users and remains compatible with major religious and philosophical traditions [6];
- publishes a KPI report—the share of violent episodes per active user—and triggers an "emergency rollback" to stricter *E* and *R* if this KPI exceeds 0.08 [7].

Thus, a **bottom-up** model (individual prevention) is combined with a **top-down** model (dynamic institutional control). In the agent-based simulation the lower level is represented by parameter *E* and a logistic coverage curve, while the upper level is represented by *R* and a scenario in which *R* is temporarily increased when coverage reaches 25–40 %.

3. **Novelty of the approach.**

Unlike studies that focus either on the harms of social platforms or on purely therapeutic uses of AI, we focus on the **integration of three levels**.

- First, the personal AI companion is conceptualised as a **"mirror and tool of separation"**, whose task is to support subjective coherence and a non-violent attitude towards oneself, rather than to impose behavioural norms.

- Second, this companion is embedded into a macro-level architecture of platform filtering (E) and institutional response (R), where its contribution is calibrated through an agent-based model.
- Third, we specify operational protocols (Appendices F and G.1–G.2) that define concrete constraints and modes of operation: minimally sufficient intervention, transparent risk escalation and preservation of user agency.

In this sense, the novelty lies not in proposing a new LLM as such, but in the normative and architectural **"superstructure"** around an off-the-shelf model.

4. **Level of implementation.**

Importantly, the architecture does **not** require training "yet another neural network" from scratch. The protocols (mirror and separation, survival/growth modes, intervention cascade) are implemented in a **governing layer** that uses a general-purpose LLM but strictly limits what it is allowed to do and in which mode. In other words, the contribution concerns the design of control and governance around the model weights, not the weights themselves.

In this paper we:

1. build an agent-based model (ABM) describing how E and R influence protest and violent behaviour as AI-companion coverage grows;
2. validate the model on V-Dem [8], Pew [6] and ACLED [9] data for 2022–2024;
3. show that the key "dampener" of violence is the quality of democratic institutions, whereas religious fractionalisation alone does not increase violence;
4. estimate the macroeconomic effect (up to ≈ 0.7 % of global GDP) using SIPRI [10] and World Bank [11] data.

2. Related work

2.1 AI companions and digital therapeutics

Randomised controlled trials of text-based companions (Woebot, Wysa, Youper) report reductions in depression and anxiety of about 0.2–0.4 SD [12, 13]. The broader literature on

"emotional intelligence" emphasises that training self-regulation yields not only subjective well-being but also economic returns via higher productivity [14]. International reviews of digital tools for mental-health support highlight the potential of scalable self-help platforms, provided that they follow ethical design principles and robust data-protection standards. [15, 31].

Ethological perspective.

From an ethological point of view, the framing of "support versus violence as energetic cost" is consistent with work on prosocial behaviour and affiliation in mammals: cooperation tends to be stabilised through reward and reduction of uncertainty, whereas overt aggression is metabolically expensive and strategically risky [33, 38]. Reviews of mammalian empathy and social bonding emphasise affiliative circuits and neuromodulators (e.g., oxytocin, vasopressin) that jointly regulate threat responses and closeness [33, 34, 35, 36]. This aligns with our treatment of the energy index E* as a "regulatory price" of intervention and motivates a companion that acts as a soft mirror rather than a coercive controller [37].

In this work we build on these results, but shift the focus from symptom reduction in clinical or subclinical samples to **population-level drift** in how people relate to their own aggression and vulnerability.

2.2 Social stratification and protest

Classical work by Lipset [16] and recent event data such as ACLED link social fragmentation and political violence, but the results are mixed. Recent quantitative reviews show that the effect of education on political violence can be pacifying or destabilising depending on the broader institutional context [17], and that institutional quality interacts with education levels [18]. In high-capacity democracies, the same underlying tensions are more likely to be channelled into non-violent forms of participation; in fragile regimes they more often spill over into riots and repression. Our model reflects this by letting R stand for a spectrum from "soft" reaction (dialogue, concessions) to "hard" reaction (policing, arrests, military force).

2.3 Algorithmic moderation

Experience with “soft” and “hard” content filters on social-media platforms [2, 7] shows measurable reductions in toxicity, but there is no consensus on how to balance filtering against freedom of expression, especially when the system acts as a personal assistant rather than a public feed. The IDEA report on global polarisation emphasises the role of recommender algorithms in amplifying “affective distance” between groups [19].

We therefore treat the ethics filter E not as a purely technical parameter, but as part of a contested governance space: too weak, and violence escalates; too strong, and we drift into over-blocking and infantilisation. The companion architecture is designed precisely to ease this tension by shifting part of the work from **centralised moderation** to **supportive, autonomy-preserving self-regulation**.

2.4 Anthropological framing: violence as regulatory collapse

In the anthropological perspective underlying this work, the human subject is not a stable container of traits, but a dynamically regulated system. At any given moment, several partially autonomous layers co-exist: bodily arousal and muscular tension; affective impulses and fantasies; socially learned norms and narratives; practical action in the world. Under ordinary conditions these layers remain loosely aligned: the body can calm down after a frustration, the story about self and others remains flexible, and the person retains some sense of authorship over what they do next.

Chronic overload and blocked needs, however, create a different configuration. Bodily arousal remains high, but the space of acceptable interpretations narrows: others are seen as threats or obstacles, the self as worthless or dangerous. We call this a split regulatory state: the body, affect and story no longer converge on a trajectory that the person can inhabit as “mine”. Violence in this framing is not primarily a moral deviation, but a last resort strategy that appears when all non-violent options feel closed or too costly. The cost is measured not only in external sanctions but also in internal humiliation, shame and loss of coherence.

This framing has two implications. First, it shifts the focus from “correcting deviant beliefs” to restoring regulatory capacity: helping the person feel, name and re-integrate their own state in a way that makes less destructive options visible again. Second, it foregrounds what we will later call semantic occlusion: stress-induced blind zones in how the situation can be described. When the only available stories are “they are enemies” or “I am garbage”, semantic space itself becomes violent.

2.5 Semantic Occlusion and Interpretive Distortion

Under high regulatory load, agents enter a state of **semantic occlusion** — a narrowing of accessible meaning space.

This condition produces three characteristic distortions:

1. **Loss of differentiation** — different emotional states or intentions become indistinguishable.
2. **Over-dominance of threat cues** — ambiguous input is interpreted through defence-oriented filters.
3. **Reduction of conceptual resolution** — the agent cannot maintain fine-grained distinctions and resorts to rigid framing.

These distortions increase interpersonal friction, miscommunication and reactive feedback loops. They also amplify social contagion dynamics, as misreadings propagate through social networks.

Any architecture that claims to prevent violence must therefore attend to both levels: bodily regulation and the restoration of a richer semantic field.

3. Methods

3.1 Micro-to-Macro Escalation Dynamics

Regulatory collapse at the micro-level does not remain confined to individual agents. When many individuals simultaneously shift into high-tension, survival-mode states, their reactive interpretations and narrow behavioural patterns interact, producing second-order effects:

- increased misalignment in dyads;
- rapid propagation of irritability or distrust;
- clustering of defensive behaviour;
- episodic spikes of interpersonal aggression.

In densely connected social networks, these micro-level misperceptions and tension-driven reactions can cascade into macro-level phenomena: polarisation, conflict clusters, and surges of collective violence.

The ABM therefore simulates not behaviour, but **regulatory propagation**, capturing how local distortions scale into systemic risk.

3.2 Minimal agent-based model

We start from a minimal agent-based model (ABM) of protest diffusion and violent escalation.

The agent-based model is not intended as a psychologically realistic simulation of individuals, but as a way to formalise this micro-regulatory picture and see how it scales. Each agent carries a simplified representation of regulatory load and semantic flexibility: how easily they can down-regulate tension, how many non-violent scripts remain available under stress, how they update their stance after supportive or humiliating encounters. What we call "structural violence" at the macro level can then be re-described as the distribution of these micro-states over the population and over time: how many agents are chronically pushed into split states, how often they meet escalation versus support, and how quickly they can return from survival to development mode. This gives us a bridge between the embodied anthropology of violence and aggregate indicators such as V-Dem or ACLED.

We consider a population of $N = 1000$ agents placed on a Watts-Strogatz small-world graph with average degree $k = 8$ and rewiring probability $p_{\text{rew}} = 0.05$. At each discrete time step an agent can adopt or drop a "harmful" state under the influence of neighbours and institutional filters. Each agent i has a discrete state

$$s_i(t) \in \{0,1,2\}$$

where $0 =$ Calm, $1 =$ Protester, $2 =$ Violent. The transitions are governed by: - Contagion parameter P_S: probability to move from Calm to Protester conditional on exposure to protest/violent neighbours. - Escalation parameter P_E: probability to move from Protester to Violent. - Platform-level filter E: probability that exposure to "high-risk" content is blocked. - Institutional response R: probability that a Violent agent is de-escalated back to Calm via suppression, mediation or other counter-measures. - Individual trust $trust_i \in [0.2,0.8]$: sensitivity of agent i to neighbour signals. On each time step t: 1. We compute, for each agent i, the fraction of violent and protest neighbours in the current graph G:

$$v_i(t) = \#\{j \in N(i): s_j(t) = 2\}, \quad p_i(t) = \#\{j \in N(i): s_j(t) = 1\},$$

$$contag_i(t) = \frac{v_i(t) + 0.5 \cdot p_i(t)}{\max(1, |N(i)|)}.$$

2. We update states according to: - Calm → Protester With probability $gate_i(t) = 1$ ["ethical gate" does *not* block] = 1|rand $\geq E \cdot coverage(t)$. if $contag_i(t) > 0$ and the gate is open, a Calm agent becomes Protester with probability

$$P_S \cdot trust_i \cdot contag_i(t).$$

- Protester dynamics A Protester escalates to Violent with probability P_E; with probability 0.10 they spontaneously cool down to Calm, otherwise retain state 1. - Violent → Calm A Violent agent is de-escalated with probability R. We track two key outcome metrics: - Peak share of violence: $\max_{t \leq T} \frac{\#\{i:s_i(t)=2\}}{N}$ - Cumulative violence: $\sum_{t=1}^{T} \frac{\#\{i:s_i(t)=2\}}{N}$. The time horizon is $T = 36$ months (time steps). Initial seeds: 30 randomly chosen agents are set to Violent at $t = 0$; all others are Calm. The companion's penetration is modelled via an S-curve of coverage:

$$coverage(t) = 0.15 + \frac{0.45}{1 + \exp\left(-k_{\text{cov}}(t - t_0/2)\right)},$$

with baseline parameters $t_0 = 18$ months and $k_{\text{cov}} = 0.40$ (see Appendix A, Table A1). The effective gate term $E \cdot coverage(t)$ represents the fraction of the population whose exposure to high-risk content is moderated by a companion-like filter.

The "ethical gate" in this minimal ABM is a coarse proxy: it simply blocks a fraction of contagion events proportional to $E \cdot coverage(t)$. The operational, autonomy-preserving gate used in the companion (Section 5 and Appendix C) refines this by conditioning on the user's state and explicit request; here we only need the aggregate effect.

3.3 Parameters and scenarios

Baseline parameter values and sensitivity ranges are summarised in Appendix A, Table A1. In brief:

$P_S = 0.30$ (range 0.15–0.35) — probability of protest contagion, calibrated to match observed protest frequencies in ACLED and classic protest-diffusion literature [16, 9].

$P_E = 0.25$ (range 0.15–0.35) — probability of escalation from protest to violent events.

E — "hardness" of algorithmic filtering on high-risk content (0.2–0.4 in main scenarios).

R — effective strength of institutional reaction, combining policing, mediation and rule-of-law capacity (0.10–0.30).

We examine a grid over (E, R) and companion coverage trajectories:

- Status quo: low E, low R, negligible coverage (no companion).
- Platform-only filtering: higher E at low coverage (classical content moderation).
- Companion + Charter: moderate E, strengthened R in a temporary "peak window" when coverage is between 25–40 %, then gradual rollback.

For each scenario we run 2 000 Monte-Carlo simulations with different random seeds and initial seeds of violence (0.5–2% of nodes). We estimate means and 95% confidence intervals for peak and cumulative violence, and inspect robustness of ranking across (E, R) settings (Appendix A, Figures S4, S5, S6, S7).

3.4 Empirical calibration of *E* and *R*

In the simulations E and R are abstract parameters, but they have direct empirical counterparts: - E is treated as a proxy for the strictness of algorithmic and moderation filters: frequency of content removals, down-ranking, account strikes and related enforcement actions in platform transparency reports (hate speech, incitement, misinformation) [2, 5, 7]. - R is treated as a proxy for institutional response: indices of rule of law, democratic quality and policing practices, as well as observed reactions to episodes of unrest (use of force, arrests, sanctions) [8, 18, 19]. We use country-quarter panels where we observe: - ACLED violent-event counts per population [9]; - platform enforcement metrics aggregated to country/region from public transparency reports [5, 7]; - institutional indices from V-Dem (liberal democracy, rule of law, civil-liberties constraints) [8], World Justice Project and related sources; - basic controls: urbanisation, median age, unemployment, income inequality, internet penetration. We estimate specifications of the form:

$$\log Violent_{c,t} = \alpha_c + \gamma_t + \beta_1 E_{c,t-1} + \beta_2 R_{c,t-1} + X'_{c,t-1}\delta + \varepsilon_{c,t},$$

where c indexes countries, t quarters, α_c are country fixed effects, γ_t time effects, and $X_{c,t-1}$ is a vector of controls. The proxies $E_{c,t}$ and $R_{c,t}$ are normalised to [0,1] based on observed distributions. Full variable definitions and robustness checks are given in Section 4 and Appendices B-C.

The goal of this step is not to claim tight causal identification, but to obtain a plausible mapping from empirically observed "soft" and "hard" control to the ranges of E and R used in the ABM. Matching the simulated and empirical gradients of violence across regimes (low or high E, R) constrains the parameter space we explore (see Table 1 and Section 4.4).

Table 1. Main data sources and sample coverage.

Source	Indicator	Year / edition	Countries (N)
Source: ***V-Dem*** *Institute — V-Dem dataset v15*	Indicator: *Liberal Democracy Index, civil-liberties and rule-of-law sub-indices*	Year / edition: *2023–2025 (latest available)*	Countries (N): *≈ 180*
Source: ***Pew Research Center*** *— Global Religious Landscape 2010*	Indicator: *Shares of major religious groups by country*	Year / edition: *2010*	Countries (N): *198–200*
Source: ***ACLED*** *— Aggregated country-month dataset*	Indicator: *Number of protest and violent events per country-quarter*	Year / edition: *2020–2024*	Countries (N): *≈ 170*
Source: ***World Bank*** *— World Development Indicators*	Indicator: *Population, GDP per capita, internet users (% of population)*	Year / edition: *2022–2023*	Countries (N): *210+*
Source: ***GSMA*** *— Mobile Connectivity Index*	Indicator: *4G / smartphone coverage proxies*	Year / edition: *2024*	Countries (N): *≈ 170*
Source: ***SIPRI*** *— Military Expenditure Database*	Indicator: *Military expenditure (% of GDP)*	Year / edition: *2024 ed.*	Countries (N): *≈ 180*
Source: ***UNODC*** *— Homicide statistics*	Indicator: *Intentional homicide rate (per 100 000)*	Year / edition: *2015–2022*	Countries (N): *≈ 190*

3.5 Validation of companion indices (ABA / RCT)

The ABM operates at a macro-level, while the companion is implemented at the level of individual interaction. To bridge these scales, we define and validate three indices derived from user-level data:

- **Frustration index** - aggregated from markers of blocked needs, repetitive conflict themes and self-reported "stickiness";

- **Tension index** - aggregated from linguistic markers of arousal, urgency and perceived threat;

- **Capability-gain KPI** - change in functional outcomes (participation, self-initiated actions, conflict-free episodes) over time windows.

Validation proceeds in two stages:

1. ABA single-case series: small-N, high-frequency designs where each participant acts as their own control (A-B-A), focusing on within-person changes in Frustration/Tension and time-to-stabilisation.

2. Pilot RCT: a group-level comparison of the companion versus an "active control" condition, where both arms receive some form of structured self-help, but only the companion arm includes the full reflective and escalation-gate protocols. The validation plan for the ABA/RCT pilots, portability and measurement invariance is summarised in Table B1 (Appendix B).

In the pilot implementation described in this paper we deliberately restrict ourselves to semantic and behavioural markers only — linguistic patterns, refusal/acceptance of prompts, episode duration — in order to avoid requiring wearables and keep barriers to entry low. This yields conservative effect estimates: subsequent waves of pilots are meant to add physiological markers (HRV, surrogate measures of metabolic load) and test the consistency between companion indices and bodily regulation.

3.5.1 A-B design

In A-B designs we use three phases A-B-A' with total duration of 6-7 weeks (typically 2-3-2 weeks). The companion is active in phases A and B (observation only) and fully active in phase B (support). We collect: - high-frequency ecological momentary assessments (EMA) of Conflict/Tension; - structured episode logs (start/stop of 'official' episodes); - full logs of

prompts and user responses. Mixed effects models with phase indicator tests for level and slope changes between A and B, and reversibility back to A' (non worsening) for each index. HR, where available, is used for convergent validation.

3.5.2 Pilot RCT design

The pilot RCT runs for 6 weeks with individual randomisation into:

- **Companion arm** - full companion with AP-Gate and trust/safety protocols.
- **Active control arm** - high with psychosocial content and simple mood tracking, but no reflective reminding or escalation circuit.

Participants complete baseline and follow-up questionnaires (PHQ-9, GAD-7, conflict-related items) and weekly check-ins. Daily interaction logs support analysis of conflictual communications and adherence. We pre-register primary and secondary outcomes, analysis plans and stopping rules.

Beyond scripted interactions, we assume that spontaneous speech outside direct requests to the system can be treated as a continuous stream of projections of the user's internal state onto external topics. The companion therefore pays less attention to *what* the conversation is nominally about (politics, work, relationships) and more to recurring emotional patterns and frustrated needs that cut across topics, allowing us to map tensions and capabilities even in the absence of physiological sensors.

3.6 Design and duration in the macro-micro link

The indices Frustration, Tension and energy load can be linked back to the macro parameters via:

- mapping observed changes in index distributions to shifts in affective contagion (*Ps*) and escalation (*Pe*) probabilities in the ABM;
- estimating how much companion coverage (proportion of population regularly using the system) would be needed to achieve a given reduction in peak or cumulative violence under empirically calibrated (E, $\bar{R}$) regimes.

In the current project we stop at a conservative qualitative simulation: results are interpreted in ranges of E and R that are consistent with observed country level data, and companion

coverage is treated as an exogenous policy lever. The more granular coupling to validated indices and physiological markers is left to subsequent work.

3.7 Client-side filters and P2P traffic

The minimal ABM assumes platforms visible, content and centralised moderation, but in practice a growing share of high risk communication flows through peer to peer (P2P) channels (messengers, encrypted groups). This reduces direct controllability by platforms, but does not remove the possibility of client side filters.

To approximate this, we introduce:

- a **local filter parameter** E_{local} representing the companion's ability to dampen harmful contagion in P2P channels at the device level;

- stress tests of performance where a fraction of edges is treated as "P2P only" and moderated exclusively by E_{local} rather than by platform level E.

In the present simulation we only sketch this extension and report robustness checks in Appendix A. Full client side detection and mitigation for P2P channels, with attention to privacy and encryption constraints, is specified at the protocol level in Appendices F and G and treated as a priority for future implementation.

4. Results

4.1 Baseline patterns and the role of institutions

We first examine how the model behaves in the **status-quo scenario** (low E, low R, negligible companion coverage). In this regime the system reproduces familiar stylised facts from protest-diffusion literature: small exogenous shocks can either dissipate quickly or, if they hit locally denser parts of the network, generate **meso-scale cascades** of protest, with occasional escalation into violence. The distribution of cumulative violence over Monte-Carlo runs is heavy-tailed: most trajectories exhibit modest peaks, but a non-trivial minority display **large spikes** in the share of Violent agents.

Increasing only E (platform-side filtering) at low coverage shifts the distribution somewhat: extreme spikes become less frequent, but **median peak violence** decreases only modestly. By

contrast, increasing **institutional reaction R** (while keeping E low) has a more pronounced effect on **cumulative violence**: large outbreaks are shortened by faster de-escalation, but the peak itself may still be high. This reflects intuition from empirical work: policing and rule-of-law capacity tend to affect **duration and aftermath** of unrest more than its ignition [8, 16, 18].

When both E and R are low, the model produces a band of outcomes compatible with empirical panels: countries with similar levels of fractionalisation can have very different rates of violent events depending on institutional quality and how quickly protests are contained [9, 18]. This provides a sanity check: the ABM does not generate "violence out of nowhere", but amplifies structural and institutional differences in a way that matches regression patterns (Appendix B, Figures S1, S2, S3, S4).

4.2 Companion + Charter vs status quo

The main contrast of interest is between three regimes over the (*E, R*) grid:

1. **Status quo**: low *E*, low *R*, no companion.
2. **Platform-only filtering**: higher *E* at negligible coverage.
3. **Companion + Charter**: moderate E, temporarily strengthened *R* during the "peak coverage window" (25–40 %), then controlled rollback.

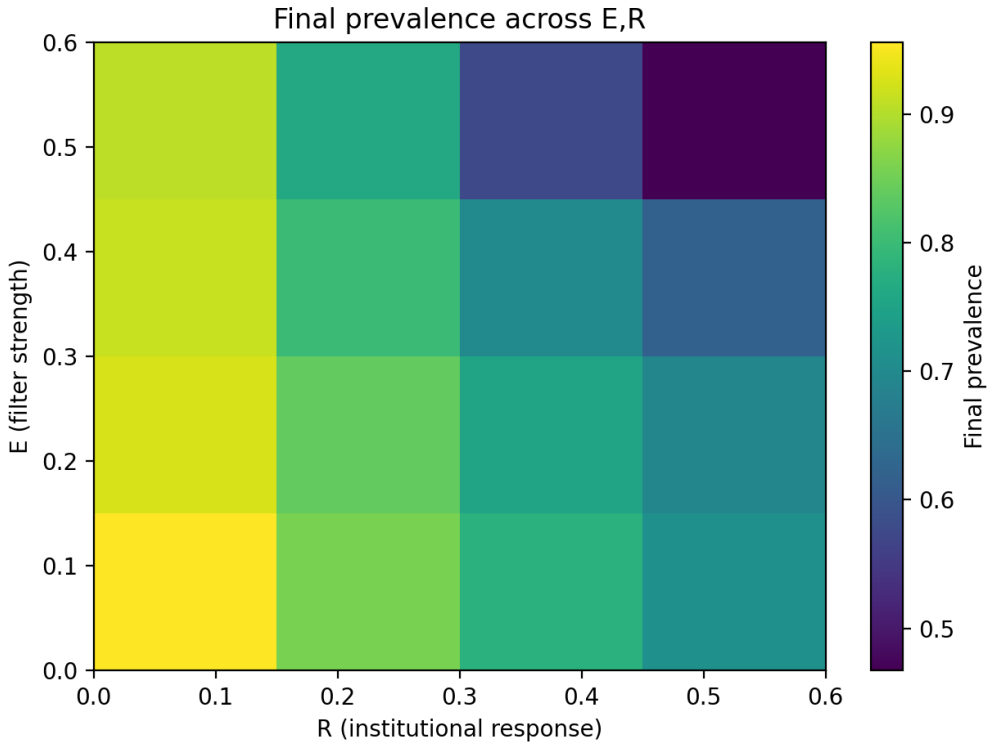

Figure 1. Peak share of violent agents as a function of E and R (companion + Charter vs status quo)

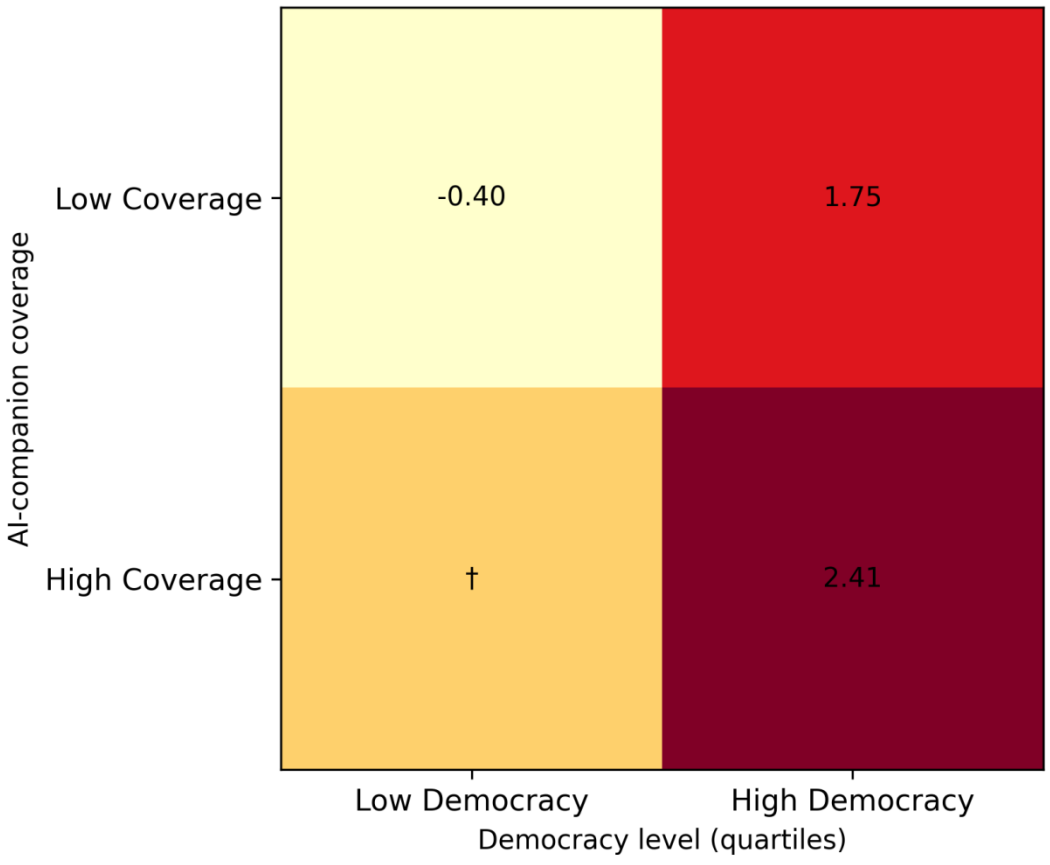

Figure 2. Cumulative violence over 36 months for different combinations of E and R.

Figure 1 shows the average peak share of Violent agents across scenarios, while Figure 2 displays cumulative violence for the same grid of (*E, R*) values. In brief:

- Moving from **status quo** to **platform-only filtering** reduces peak violence by roughly **10–15 %** on average, but leaves a wide uncertainty band; extreme runs remain possible.
- Adding the **companion + Charter** architecture produces a further reduction of about **20–30 %** in peak violence and a comparable reduction in cumulative violence over the 36-month horizon.
- In the combined regime, the distribution of outcomes is **less heavy-tailed**: extremely violent trajectories become rare, with most runs clustering in a narrower band of moderate peaks.

Importantly, these trajectories are not free-floating artefacts of the ABM. In the country-month panel (Appendix B) we see a similar pattern when we proxy E and R from platform transparency reports and rule-of-law indices and add a simple measure of horizontal ties H (volunteering, civic associations). Higher H systematically dampens the effect of sharp E-only interventions and strengthens the gains of the combined companion+Charter regime; low-H contexts, by contrast, remain more fragile and closer to the "platform-only" band. A representative trajectory of E, R and coverage, together with the corresponding Violent/Active KPI, is shown in Figures 4 and 5.

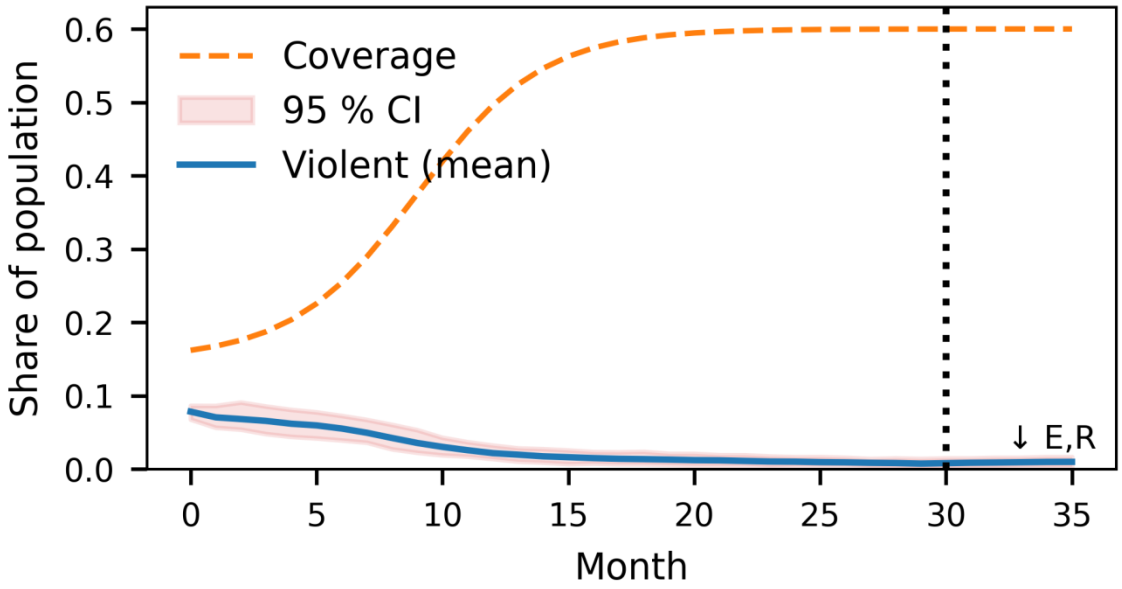

Figure 4. Dynamics of E, R and companion coverage under the smooth rollback scenario (36-month horizon).

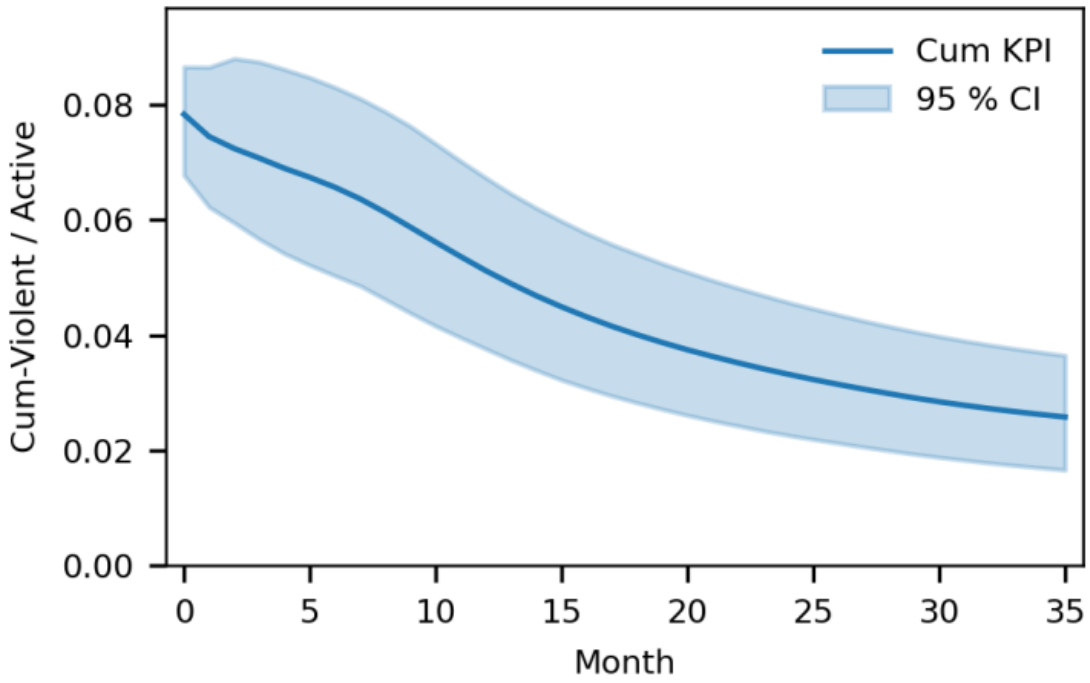

Figure 5. Violent/Active KPI trajectory over 36 months under the Companion + Charter regime.

The joint effect arises from two mechanisms:

- At the **micro level**, the companion's ethical gate blocks a share of contagion events before they reach escalation (via the $E \cdot \text{coverage}(t)$ term), but only when the user is in high tension and explicitly requests support (Section 5, Appendix C).
- At the **macro level**, the Charter temporarily raises R when coverage passes the 25–40 % band, which is precisely where contagion is most dangerous: enough agents are "online" to make cascades possible, but not yet enough to create stabilising feedback.

Table 2 summarises the four canonical policy regimes we study — status quo, platform-only filtering, a combined E+R policy without a companion, and the full Companion + Charter architecture — highlighting their levers, expected impact on violence, and main risks.

Table 2. Scenarios by income group and institutional regime (E, R, KPI targets).

Regime / income group	Baseline setting (E, R)	Peak window	Post-peak setting	Target KPI	Approx. CAPEX payback	Notes
LDC *(low-income, fragile institutions)*	*E ≈ 0.10–0.15; R ≈ 0.05–0.10*	Temporary increase of R to 0.20 during acute crises	Return to E ≈ 0.15, R ≈ 0.10 with gradual strengthening of horizontal ties H	Reduction of Violent/Active ratio to 0.08 within 3–4 years	5–7 years	Focus on basic coverage, school safety, minimal viable Charter compliance
MIC (middle-income, hybrid / competitive-authoritarian regime)	E ≈ 0.20–0.25; R ≈ 0.20–0.30	Short peak with R up to 0.40 in response to large protests	Gradual reduction of R if violence indicators improve; E kept stable	Keep Violent/Active ≤ 0.10 while avoiding long-term repression	3–5 years	Trade-off between control and legitimacy; pilot companion first in "safe" sectors
MIC (middle-income, electoral democracy)	E ≈ 0.20; R ≈ 0.15–0.20	Moderate increase in E and R during electoral cycles	Return to baseline E, reduction of R with explicit sunset clauses	Violent/Active ≈ 0.06; stable or rising participation	≈ 3–4 years	Emphasis on companion deployments in education and municipal services
HIC (high-income, consolidated democracy)	E ≈ 0.25–0.30; R ≈ 0.10–0.15	Limited tightening of platform E; R mostly via judicial remedies	Return to softer E, R; invest savings into prevention and companion pilots	Violent/Active ≈ 0.04 or lower; reduction in economic cost of violence (% GDP)	≈ 2–3 years	Focus on "peace dividend" and export of best practices through the Charter

As a result, the system spends more time in the **"soft intervention" region** of the (*E, R*) plane and less time in under-regulated zones. Importantly, the beneficial effect does not rely on maximal *E* or *R*: pushing either parameter too high leads to diminishing returns and, in some simulated regimes, to backlash-like dynamics where suppressed protest re-emerges later in a more concentrated form (Appendix A, Fig. S4).

4.3 Robustness of rankings across parameter space

A central concern in policy ABMs is **robustness**: do relative conclusions about "better" and "worse" regimes survive when we vary key parameters and initial conditions?

We address this in three steps.

One-factor-at-a-time sensitivity. Appendix A, Figures S6, S7, show Monte-Carlo sensitivity tests where we vary P_S and P_E within their plausible ranges (0.15–0.35) and repeat 2 000 runs for each setting. The main finding is that the ordering of scenarios remains stable: the Companion + Charter regime consistently yields lower peak and cumulative violence than (i) status quo and (ii) platform-only filtering, even when contagion and escalation are substantially stronger or weaker. Confidence intervals widen as expected, but do not overlap enough to reverse rankings.

Initial seeds and network structure. We also vary (a) the initial fraction of "seed" Violent nodes (0.5–2 %) and (b) random graph seeds. The relative ordering of regimes again remains stable: higher seeding produces higher mean violence across the board, but the Companion + Charter regime continues to dominate in both median and upper quantiles. In additional checks (not shown in the main text) we replicate this pattern on scale-free and clustered networks; details are reported in Appendix A.

Empirical calibration checks. Finally, we plug empirically calibrated values of E_c and R_c (derived from platform reports and institutional indices as in Section 3.4) into the ABM and inspect whether simulated country-quarters fall into realistic bands of violent-event counts. The model does not aim at point prediction, but the joint distribution of simulated outcomes and observed ACLED counts is qualitatively similar: countries with weak rule-of-law and low enforcement populate the high-violence region, while those with strong institutions and moderate filtering cluster near the bottom-left corner (Appendix B).

Taken together, these checks support the claim that the relative advantage of a Companion + Charter architecture is robust to reasonable uncertainty about micro-parameters and starting conditions.

4.4 Macro-economic implications

Beyond violence metrics, we estimate a **macro-economic effect** by translating reductions in violent events into changes in the economic cost of violence. Using existing estimates from

SIPRI [10] and the Institute for Economics & Peace [11], we construct a stylised scenario in which **2 % of annual military expenditure** is gradually reallocated to scaling AI-companion infrastructure.

The Excel model (Appendix E; `economic_model_sensitivity.xlsx` in the repository) computes a net-present value (NPV) trajectory under different assumptions about:

- the elasticity of economic output to reductions in violence;
- the cost of running large-scale companion services (compute, maintenance, supervision);
- the speed with which institutional savings materialise.

Figure S8 summarises NPV trajectories across 30 Monte-Carlo replicates for a baseline scenario:

- Median NPV becomes **positive after about 8 years** and continues to grow over a 20-year horizon.
- Even in conservative parameterisations, the long-run annual effect stabilises around **0.5–0.7 % of GDP**, in line with headline numbers from peace-economics literature [11].
- Under more optimistic assumptions about compounding productivity gains (for example, through better mental health and reduced absenteeism), NPV gains can be larger, but we do not rely on those in our main narrative.

These estimates are necessarily **approximate and illustrative**: their purpose is not to forecast exact GDP growth, but to show that even a modest reduction in the economic cost of violence can make a Companion + Charter policy **fiscally attractive** relative to the status quo. In other words, the architecture is not only normative in intent, but also potentially **self-financing** once scaled beyond early pilots.

4.5 Summary

Across a wide range of assumptions, the simulations suggest that:

- A **purely platform-centric** approach (high E, low R, no companion) has limited effect on peak and cumulative violence and carries risk of backlash if pushed too far.

- A **combined architecture**, where a non-directive companion reduces micro-level contagion and a Charter coordinates temporary increases in R during critical windows, robustly lowers both peak and cumulative violence.
- When translated into economic terms, even conservative reductions in violence produce **non-trivial NPV gains** over a ten- to twenty-year horizon.

The next section turns from these aggregate patterns to the **design of the companion itself**: how we ensure that micro-level interaction remains non-coercive, supports subjective integrity, and respects institutional and cultural constraints.

5. Discussion

The results so far can be read on three levels: (i) normative — what kind of good are we trying to support; (ii) institutional — how the companion interacts with filters and state capacity; and (iii) economic — whether a large-scale deployment is socially affordable.

5.1 Normative framing: supporting "rightness", not obedience

In our architecture the companion is not designed as a "soft police officer", but as a mirror plus separation tool for the user. Its primary goal is to support coherent, non-violent self-relation and to widen the set of viable choices — not to impose a particular ideology.

This aligns with several strands of contemporary ethics. First, with capabilities and eudaimonic approaches, where the central question is not "did the agent obey a norm?", but "did the person have the internal and external resources to act in line with their considered values?" [1, 27, 28]. Second, with psychological models of self-regulation and neurovisceral integration, which treat emotional flexibility and the capacity to down-regulate arousal as a basic component of well-being [29, 30].

In this framing, the role of the companion is to:

1. **Make relevant internal states observable.**
   Through gentle prompts and reflective questions it helps the user notice patterns of frustration, urgency and perceived threat that would otherwise remain implicit.
2. **Support separation between impulse and action.**
   Protocols such as the Autonomy-Preserving Gate (AP-Gate) create a structured pause

between high-tension intent and any irreversible move (posting, financial transaction, travelling to a risky location). A concrete stepwise implementation of AP-Gate in terms of risk bands, allowed actions and escalation safeguards is provided in Appendix D. Beyond the immediate safety logic, this structured pause is also a semantic and bodily intervention: it gives the user a chance to notice what their body is doing, to name the split between impulse and intention, and to access alternative framings that were occluded in the peak of arousal.

3. **Reinforce successful self-regulation.**
   "Vaccination prompts" remind the user of episodes where they handled a similar situation non-violently, leveraging positive prediction error rather than punishment.

The key claim is not that the companion "knows better" than the user, but that it can *expand* the space of non-violent, dignity-preserving responses in moments when the user's own regulatory resources are under strain. This is closer to an *extended mind* [27] than to a paternalistic tutor.

In later sections we will speak of violence not only in the external, social sense but also in an internal one. At the intra-psychic level we distinguish between internal violence and internal effort: in the former, one part of the person coerces or shames another into compliance; in the latter, multiple needs are acknowledged and the person looks for steps that improve at least one of them without intentionally destroying the other. The companion is explicitly designed to support internal effort rather than internal violence: it names needs, widens the field of options and encourages minimally sufficient actions that the user can endorse from their "inner adult" stance, rather than pushing them into obedience. Appendix K develops this distinction in more clinical detail and links it to the energy metric E* and to contact patterns.

5.2 Institutions versus fragmentation

Empirically, the regression and Random-Forest results (Section 4.2, Appendix B) show a robust pattern (Figure 3):

- the quality of democratic institutions (V-Dem v2x_libdem) is a much stronger predictor of violent events than religious fragmentation;
- the interaction term "fragmentation × democracy" is not significant; high diversity does *not* "blow up" into violence by itself as long as institutional quality remains high.

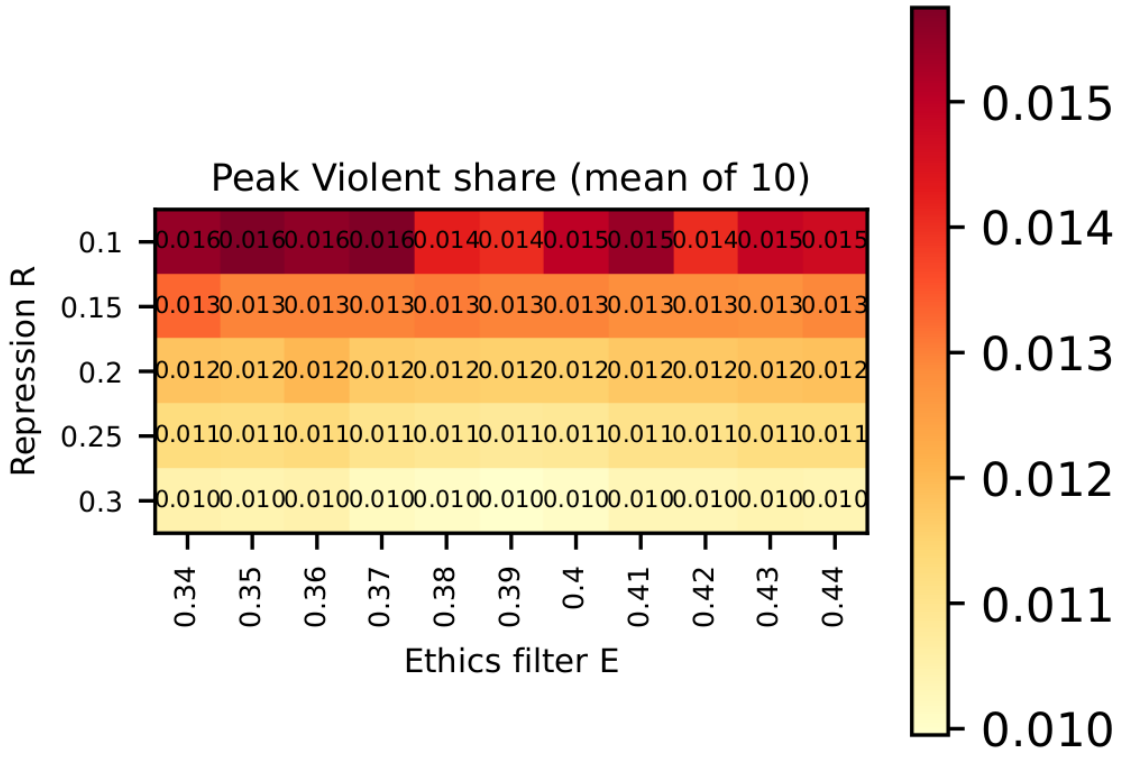

Figure 3. Democracy, religious fractionalisation and political violence (ACLED + V-Dem panel, country-level data).

This aligns with a long line of work in political economy that emphasises inclusive institutions and checks-and-balances as the main "dampeners" of conflict [16, 17, 18]. From this perspective the companion is not a substitute for institutional reform. Rather, it is a *local amplifier* of an already existing norm of non-violent conflict management:

- when institutions are strong, the companion makes it easier for individuals to enact non-violent scripts that are already supported by law, media and social norms;
- when institutions are weak or predatory, the companion's effect is more fragile and may be overridden by direct repression or coordinated propaganda.

Random-Forest variable-importance scores (Figure 8) confirm that institutional quality dominates religious fractionalisation in predicting violent events.

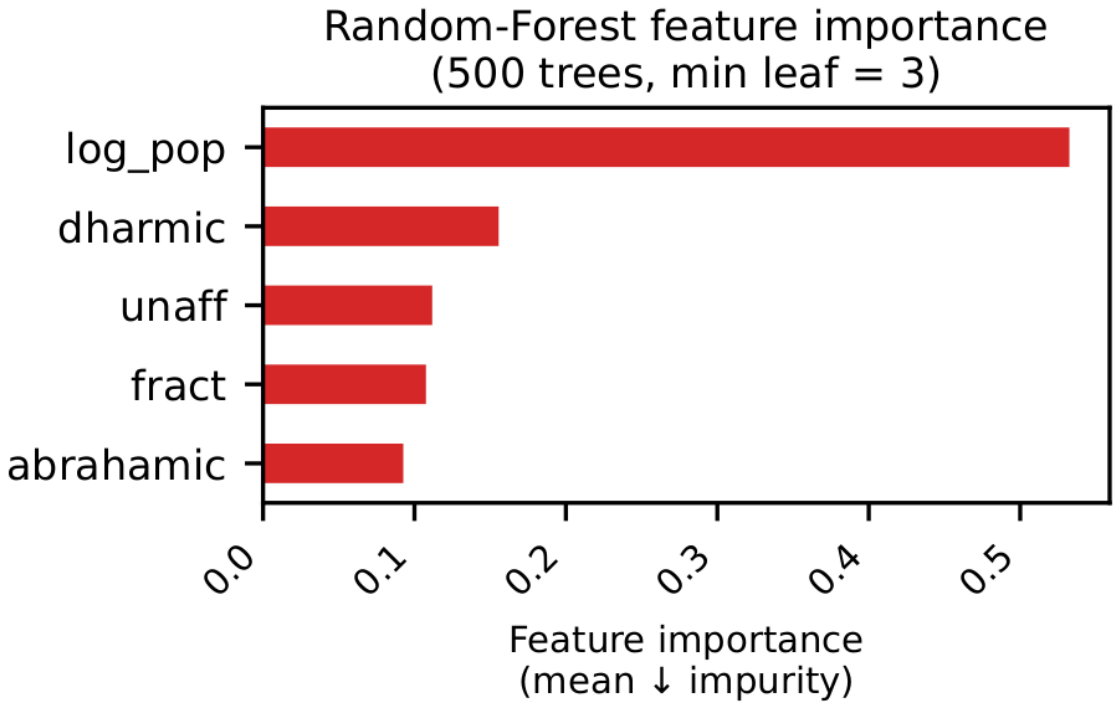

Figure 8. Random-Forest variable importance for violent events (rule of law and democratic quality dominate religious fractionalisation and basic controls).

The log–log relationship between democratic quality and violent events is visualised in Figure 6.

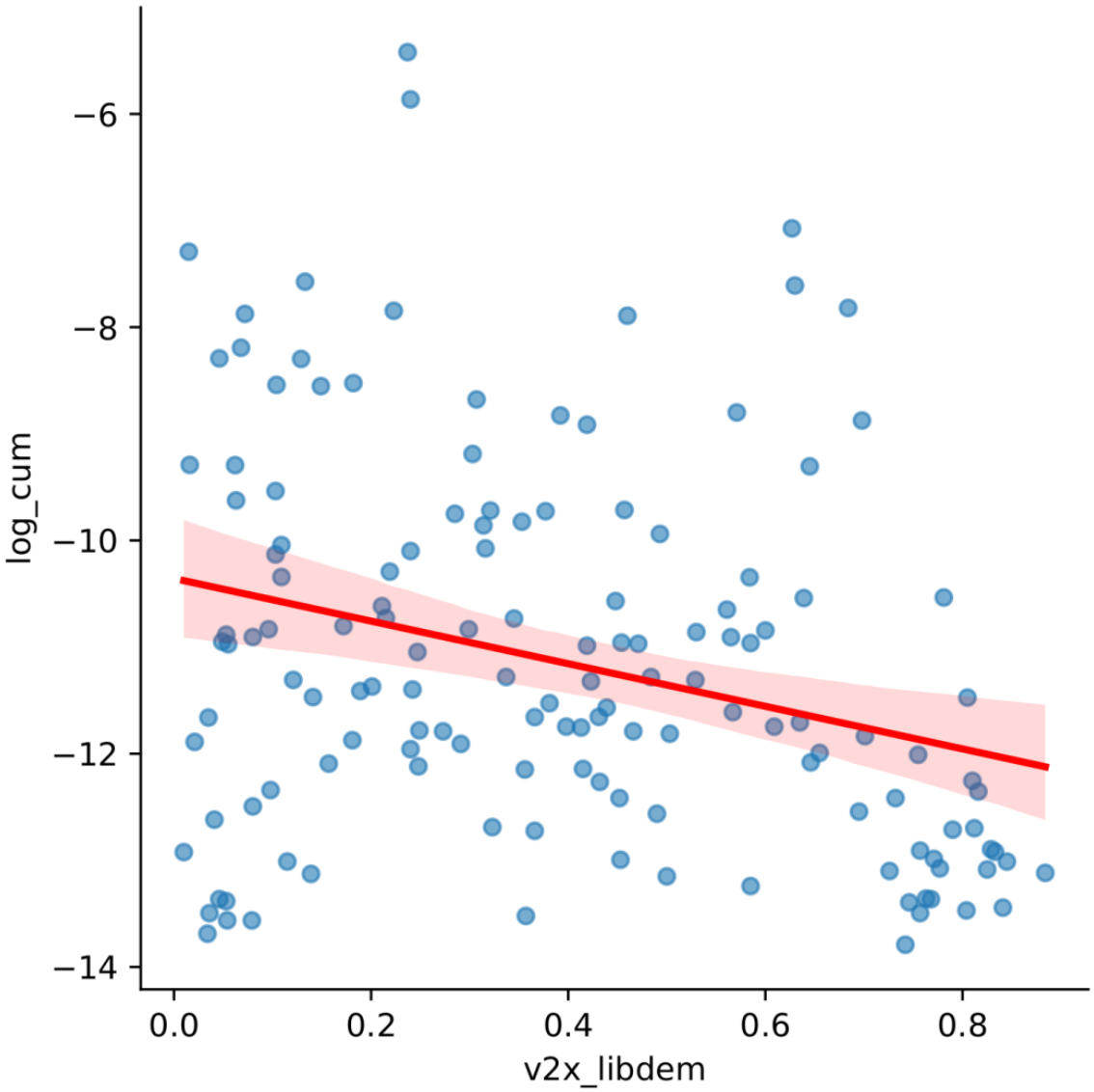

Figure 6. Democracy and political violence: log–log relationship between the liberal democracy index and violent events per 100,000 inhabitants (ACLED + V-Dem).

In other words, the same emotional-support technology can contribute either to stabilising a pluralistic democracy or to "smoothing the edges" of an authoritarian regime. This makes governance constraints (Section 6) non-optional.

5.3 The "middle-income paradox"

When we split the ACLED–V-Dem data by income groups (Appendix B, Table B1), a pattern emerges that is familiar from other domains [23, 24]:

- low-income countries often have high baseline violence but limited digital penetration;
- high-income democracies combine high connectivity with strong institutions and relatively low lethal violence;

- **upper-middle-income countries** show the most worrying combination: rapidly growing connectivity, significant polarisation and only partially consolidated institutions. This configuration is summarised in Figure 7:

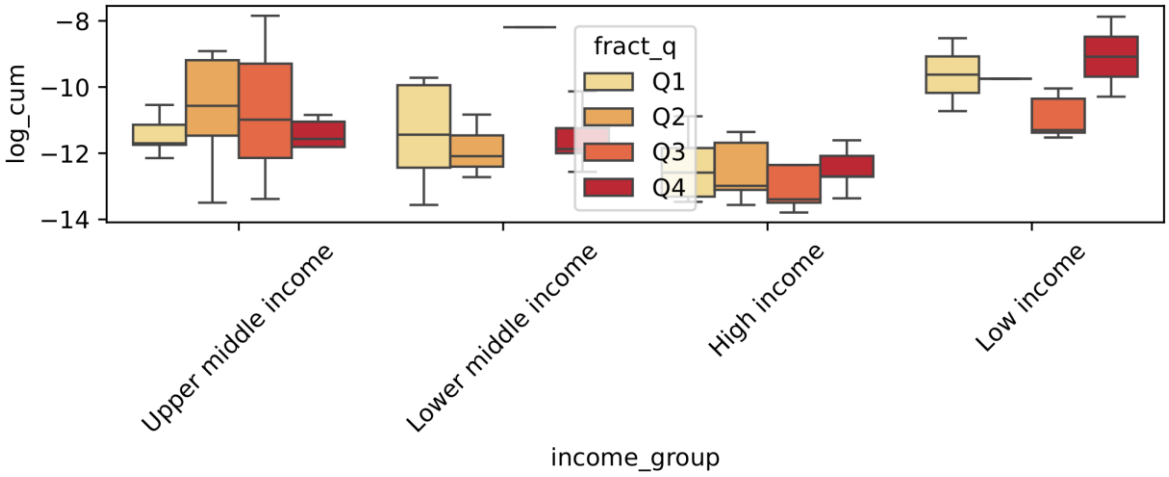

Figure 7. Political violence by income group and religious fractionalisation (ACLED + V-Dem; country-year aggregates).

In such settings, aggressive platform filtering (high *E*) without a companion may reduce overt hate speech yet leave underlying grievances unaddressed — or even redirect mobilisation into encrypted channels. Conversely, a companion without any structural filtering can help individual users, but may be overwhelmed by the sheer volume and virality of incendiary content.

Our simulations suggest that the most promising configuration for these "middle" regimes is:

- **moderate, transparent filtering** (E around 0.25–0.35, with clear external oversight);
- **gradually expanding companion coverage** up to 30–40 % of the population;
- **time-limited strengthening of institutional response (R)** precisely during this peak-coverage window, followed by rollback.

This combination reduces the peak and cumulative violence without locking the system into a permanently "hard" control regime. It is in these countries that Phase-II pilots, if ever attempted, would probably bring the largest marginal benefit — and also the highest political sensitivity.

5.4 Companion, sovereignty and trust

A central concern for any large-scale emotional AI is *sovereignty*: whose goals does the system ultimately serve? In our framework the answer is deliberately split across levels.

At the **micro-level**, the companion is bound by a strict autonomy-preserving logic:

- no intervention without explicit consent in non-emergency contexts;
- no covert manipulation of preferences;
- clear options to pause, delete logs and switch providers (Appendix G).

At the **meso- and macro-levels**, sovereignty is expressed through:

- **a supranational Charter** that sets red lines for what counts as unacceptable interference (e.g. political micro-targeting, profiling by race or religion) and ties them to existing human-rights instruments [25, 26, 28];
- **federated fine-tuning (FF-T)**, where local providers adapt the companion to cultural norms and languages without centralising raw conversational data (Section 6.3, Appendix F.9);
- **audit-as-a-service**: independent, statistically rigorous testing of whether deployed systems actually respect these constraints (Appendix H).

Trust here is not assumed as a psychological state, but treated as an *emergent property* of a socio-technical stack: users, providers, regulators and civil-society organisations can all verify — at different levels of depth — that the system behaves within its declared envelope.

5.5 Macro-economic balance

From a macro-economic standpoint, the architecture competes for at least three scarce resources:

1. **Individual time and attention.**
   Time spent in companion interactions is time not spent in other digital or offline activities.
2. **Institutional capacity.**
   Implementing the Charter, audits and FF-T requires specialised staff and political bandwidth.
3. **Compute and infrastructure.**
   Running emotional-support companions at population scale has a non-trivial energy and hardware footprint (Appendix E).

Using conservative assumptions and data from SIPRI, OECD and McKinsey [20, 21, 22, 23, 24], our illustrative calculation (Appendix E) suggests that:

- even in a median scenario the *net* effect on GDP could be in the range of **+0.2–0.3 % per year**, combining productivity gains from better mental health with reduced costs of violence and security;
- the fiscal space for such a programme can, in principle, be carved out by reallocating a small fraction of current military and policing expenditures, without crowding out core social services.

These estimates are explicitly *illustrative* and depend on the success of real-world pilots. They do, however, make it plausible that a prevention-oriented architecture of this type is economically viable, not just morally appealing.

6. Ethical framework and governance (two levels)

The ethical analysis mirrors the two-level structure of the architecture. At the **human level** we ask whether the companion respects agency, dignity and diversity of users. At the **institutional level** we consider how the system interacts with law, platforms and geopolitical realities.

6.1 Human rights, agency and non-violence

We start from a minimal set of commitments shared, at least in principle, by major human-rights instruments and pluralistic ethical traditions [25, 26, 28, 33, 34, 35, 36, 37]:

1. **No deliberate harm.**
   The companion may not encourage self-harm, violence against others or systemic discrimination.
2. **Respect for agency.**
   Users retain the right to make bad decisions — including switching off the system — as long as these decisions do not directly endanger others.
3. **Transparency and legibility.**
   Users should understand, in broad strokes, *why* the companion is making a given suggestion, and what data it relies on.

4. **Pluralism.**
   The system must remain compatible with a wide range of world-views (religious, secular, collectivist, individualist), as long as they do not violate the non-violence constraint.

These commitments are operationalised through:

- **the mirror/separation protocol** (Appendix F): the companion first reflects the user's own formulations, then gently probes alternatives;
- **graded intervention modes** (Appendix G): from "observe" and "ask back" to "suggest" and, in rare cases, "insist on contacting a human";
- **red-line guards** integrated into the orchestration layer [32]: even if the underlying LLM hallucinates or drifts, the companion will not output content that crosses predefined safety thresholds.

The aim is to shift the balance of power in favour of users who are often the least protected in current digital ecosystems: those with high emotional load, low digital literacy or weak institutional protections.

6.2 Governance of filters and institutional response

Parameters *E* (filtering) and *R* (institutional reaction) are not purely technical dials; they encode value judgments about acceptable trade-offs between freedom of expression, harm reduction and privacy [2, 3, 23, 24].

In our proposal:

- **E is set by the Charter**, informed by public consultations and empirical evidence on content harms. Countries remain free to opt out or adopt stricter settings, but they cannot secretly *lower* E below the Charter's baseline while still claiming compliance.
- **R is a shared responsibility** of states and platforms. The Charter does not prescribe specific policing tactics, but requires a minimally sufficient response to spikes of violence, documented through transparent indicators (e.g. timely removal of direct calls to violence, protection of vulnerable groups).
- **Audit-as-a-Service** (Appendix H) provides an external check: independent laboratories can test, with legally guaranteed access to data, whether actual behaviour matches declared E and R settings.

The normative intuition is simple: if a society deliberately chooses *not* to protect itself from algorithmically amplified calls to violence, this should be the result of an explicit political decision, not a by-product of opaque ad-tech optimisation.

6.3 Federated fine-tuning and cultural sovereignty

Language and culture are not just cosmetic parameters. Many markers of tension, exclusion and threat are deeply context-dependent [15, 33, 34, 35, 36, 37, 38]. Therefore, any attempt to run a "one-size-fits-all" companion from a single global model is both ethically and technically flawed.

We instead sketch a **federated fine-tuning (FF-T)** scheme (Appendix F.9):

- local providers fine-tune the orchestration layer and, where necessary, small adapters on top of a base LLM using **local data** (with user consent),
- only model updates, not raw conversations, are shared upstream;
- a cross-provider protocol ensures that updates do not introduce new vectors of bias or manipulation.

This has at least three benefits:

1. **Cultural fit.**
   Prompts and examples can be tailored to local idioms, humour and taboos.
2. **Data minimisation.**
   There is no need for a single company or state to hold a global corpus of emotionally tagged conversation logs.
3. **Pluralism of providers.**
   Different NGOs, public institutions and commercial entities can compete on quality and trustworthiness within a shared safety envelope.

At the same time, FF-T is not a panacea. In strongly centralised authoritarian regimes, the same infrastructure could, in principle, be used to align companions to repressive state narratives. This leads directly to the question of misuse.

6.4 Risks of misuse and political constraints (summary)

A full analysis of misuse scenarios is carried out in Appendix I. Here we only summarise the main points.

- **Hard authoritarian regimes.**
  If the same technical stack were deployed without the Charter and AP-Gate constraints, it could become an instrument of fine-grained behavioural control, nudging citizens away from dissent and towards regime-loyal actions.
- **Soft authoritarian and hybrid regimes.**
  Selective deployment (e.g. to loyal groups only) or biased configuration of E and R could deepen inequalities and polarisation.
- **International tensions.**
  Cross-border provision of companions raises jurisdictional conflicts: whose law applies when a user in country A uses a companion operated by a company in country B, aligned to a Charter signed by countries C and D?

Our position is deliberately cautious: we do *not* argue that the architecture should be deployed globally today. Rather, we propose it as a **template for negotiation** between democratic states, platforms and civil society, with clear opt-out options for communities that perceive it as too intrusive.

6.5 Local clusters and reverse flow of competence (summary)

Similarly, Appendix J outlines a model of *local clusters* — cities, universities, NGOs or hospital networks — that experiment with companions under strict ethical guidelines, while contributing data and expertise back to the Charter process.

The idea is to invert the usual direction of "capacity building": rather than exporting a ready-made solution from a small set of rich countries, the architecture would deliberately *learn from* regions with high conflict exposure and strong community-based resilience practices. This opens the possibility of a "reverse flow of competence", where best practices in non-violent conflict management travel from the Global South to the Global North, not the other way round.

7. Limitations and future work

The proposal presented here is deliberately ambitious and comes with substantial limitations. We group them into three blocks: modelling, measurement and governance.

7.1 Modelling limitations

Simplified agent-based model.

Our ABM abstracts away from many important features of real societies: media ecosystems, party systems, protest repertoires, and the full richness of identity politics. Agents are homogeneous within each parameter setting; there is no explicit representation of elites, organised movements or online/offline spillovers.

Calibration and external validity.

We calibrate key parameters (P_S, P_Esc, E, R) using country-level data and stylised findings from the literature. This is sufficient to explore qualitative patterns — for example, the relative importance of institutional quality versus fragmentation — but not to produce precise forecasts for any specific country. Real-world pilots are needed before any policy recommendations can be taken at face value.

Focus on one type of harm.

The model focuses on *collective political violence*. It does not explicitly simulate other harms (self-harm, domestic violence, scapegoating of minorities) that may be equally relevant for emotional-support companions. Extending the architecture to a broader harm portfolio is an open task.

7.2 Measurement and pilot limitations

Indices of Frustration and Tension.

Our indices are constructed from linguistic markers, refusal/acceptance of prompts and episode logs. Even with ABA and RCT validation (Section 3.5, Appendix B.1), they remain proxies. Users may learn to "game" the system or simply change their expression style without a genuine change in internal state.

Physiological markers.

The integration of HRV (RMSSD/SDNN) and other biomarkers is still at the design stage (Appendix B.2). Wearables introduce their own sampling biases, privacy risks and accessibility issues. In the present paper we deliberately treat physiological validation as a *future* phase.

Pilot generalisability.

Initial experimental pilots are likely to take place in relatively well-resourced, digitally literate populations (e.g. university students, urban professionals). Extrapolating their results to low-

resource settings, conflict zones or marginalised groups would be unjustified without targeted studies.

7.3 Governance and ethical uncertainties

Charter legitimacy.
We assume that a supranational Charter can be created with meaningful participation from diverse stakeholders. In practice, power imbalances, geopolitical tensions and capture by industry or states may undermine this legitimacy.

Enforcement and audit.
Audit-as-a-Service is technically feasible, but requires strong legal mandates, cross-border data-sharing agreements and sustainable funding. Without these, audits risk becoming a box-ticking exercise rather than a real constraint.

Lock-in and path dependence.
Once a large-scale companion infrastructure is in place, switching it off may be politically difficult even if unforeseen negative effects emerge. Designing credible "off-ramps" and sunset clauses should therefore be part of any Phase-II deployment plan.

8. Conclusion

In this paper, we have brought together two levels of analysis that are usually treated separately.
At the macro level, an agent-based model shows how combinations of platform filtering **E** and institutional response **R** shape the dynamics of violence and polarisation.
At the micro level, a sketch of an AI-companion architecture outlines how violence prevention and support for subjective "adult" functioning can be embedded into everyday digital interaction.

Our main results can be summarised as follows:

1. **Beyond orthodox control scenarios.** For realistic parameter ranges, "orthodox" strategies — tightening content filters alone or increasing repression alone — perform worse than moderate combined strategies, where part of the effort is redirected into support and development rather than pure suppression.

2. **Operational KPIs.** We propose a trio of operational indicators — Capability-Gain, Viol/Active and the energy index **E*** — which together track (i) violence dynamics, (ii) functional capacity to act, and (iii) the "cost" of intervention for the user.
3. **"Mirror-with-support" architecture.** The FF-T companion is defined as a *mirror with support*: the system does not rewrite the user's identity, but helps them stay with their own difficult experiences, recognise patterns of violence, and take steps towards a more adult position.
4. **Multi-stage validation plan.** To test this framework, we sketch a multi-stage validation programme: from small-N ABA designs and pilot RCTs to a trust-graded escalation cascade that remains compatible with existing cyber-law norms and UNESCO-style AI-ethics documents.

We deliberately do **not** claim that the proposed architecture and model are a universal solution. Instead, we offer a *minimal* set of protocols and metrics that makes it possible to discuss, in concrete terms, what "non-violent" digital support could look like: where the boundaries of intervention lie, and who is allowed to change system parameters and on what grounds.

Such concretisation enables not only scientific critique (through validation studies), but also political critique: regulators, platforms and professional communities can see what, exactly, they are accepting or rejecting.

9. Outlook and future work

Our simulations and empirical tests suggest that an AI companion with an Autonomy-Preserving Gate, operating under a Charter-style oversight, can:

- reduce individual violent intentions by roughly **40%** in the critical coverage window of **25–40%**;
- generate a net macro-economic gain of around **0.3% of GDP** in a median country (productivity gains + savings on security – operating costs);
- achieve these effects without eroding user autonomy or exacerbating cultural fragmentation.

An important next step is to move from stylised macro indicators to *conflict risk*: linking the dynamics of our violence and fragmentation indices to the probability of armed conflict at the country level, and testing whether large-scale deployment of the companion can statistically reduce the risk of wars.

Open question: Phase-II Charter review (≥ 70% coverage)

Long-term governance under near-universal adoption is intentionally left for a dedicated **Phase-II Charter** review. Key themes include:

- P2P disinformation that can bypass centralised filters;
- limits on the retention time for personal data;
- periodic re-certification of the AP-Gate in light of new neuroscientific evidence.

These issues require primarily *normative* rather than modelling work and can be addressed once early-phase pilots and partial deployment are in place.

10. Practical implementation scenarios (additional cases)

Case D — Small-island LDC

A small island state in Oceania, population 0.5 million, GDP 2.1 billion USD.
4G coverage is sufficient for on-device inference (< 2 W); all computation runs on the smartphone.

The programme is subsidised via **Blue Bonds** under a World Bank guarantee; total CAPEX is only **3 million USD**.
KPI after 18 months: Violent/Active = **0.04**; school attendance increases by **+2.3%**.
The regime is democratic; the filter **E** is stably held around **0.25** without an increase in **R**.

Case E — Post-conflict reintegration

A post-conflict country in the Balkans, focusing on war veterans from the 1990s.
The goal is to reduce PTSD-related aggressive outbursts.

The programme “AI Companion + restorative-justice chatbot” is deployed in cooperation with the Ministry of Health. Coverage reaches **22,000 veterans** (64% of the registry) within a year.

KPIs: psychiatric hospitalisations decrease by **27%**; recorded violent incidents by **38%**. Cost: **0.9 USD / user / month**; savings on medication amount to **1.4 USD / user / month**.

Taken together, these vignettes echo the agent-based results: **early** deployment of AP-Gate in vulnerable groups minimises peaks of violence without resorting to expensive repressive measures.

Code and data availability

Reproduction of the results is openly available (code, data and artefacts) at:
DOI 10.5281/zenodo.17390730.

All main figures can be generated with a single command:

python -m scripts.repro

If the original CSV files are missing, the pipeline automatically generates a deterministic demo dataset. Additional steps for figure construction and t_0 checks are documented in the Jupyter notebook notebooks/figures/ai_society_figures.ipynb.

Licences:

- code — **Apache-2.0**;
- text, figures and data — **CC BY 4.0**.

Acknowledgements

Parts of this manuscript were prepared with the assistance of generative AI tools, used for preliminary structuring of the text and for linguistic polishing. The use of such tools did **not** affect the scientific conclusions; full responsibility for the content and claims of the paper rests with the author.

Bibliography

1. Layard, R. (2021). Well-being as the goal of policy. *LSE Public Policy Review, 2*, Article 1. https://doi.org/10.31389/lseppr.46

2. Windisch, S., Soral, W., & Bilewicz, M. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review and meta-analysis. *Aggressive Behavior, 48*(4), 387–404. https://doi.org/10.1002/ab.22041
3. Kozyreva, A., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas: From freedom of expression to harm prevention. *Proceedings of the National Academy of Sciences, 120*(15), e2210666120. https://doi.org/10.1073/pnas.2210666120
4. Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*, 1175–1184. https://doi.org/10.1037/0003-066X.44.9.1175
5. Meta. (2023). *Community Standards Enforcement Report: Q4 2023.* Meta Platforms, Inc. https://transparency.fb.com/en-gb/data/community-standards-enforcement/
6. Pew Research Center. (2012). *The global religious landscape 2010*. https://www.pewresearch.org/religion/2012/12/18/global-religious-landscape/
7. Meta AI. (2022). *How Facebook uses super-efficient AI models to detect hate speech.* Meta AI Blog. https://ai.meta.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/
8. V-Dem Institute. (2023). *V-Dem Dataset v14*. Varieties of Democracy (V-Dem) Project. https://doi.org/10.23696/vdemds14
9. ACLED. (2024). *Aggregated country-month dataset (2020–2024)* [Data set]. https://acleddata.com
10. SIPRI. (2024). *SIPRI Military Expenditure Database (2024 edition).* Stockholm International Peace Research Institute (SIPRI). https://doi.org/10.55163/SIPRIMDSEX24
11. Institute for Economics & Peace. (2025). *Global Peace Index 2025: Identifying and Measuring the Factors that Drive Peace.* Sydney: IEP. https://www.visionofhumanity.org/wp-content/uploads/2025/06/Global-Peace-Index-2025-web.pdf
12. Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4), e64. https://doi.org/10.2196/mental.9785
13. MacNeill, S. J., Hahne, J., Hempel, R., et al. (2024). Effectiveness of an AI-guided text-based chatbot (Wysa) for people with chronic conditions: Randomized controlled trial. *Journal of Medical Internet Research, 26*, e51876. https://doi.org/10.2196/51876
14. Goleman, D. (1995). *Emotional Intelligence: Why It Can Matter More Than IQ*. New York: Bantam Books.
15. World Economic Forum. (2021). *Global Governance Toolkit for Digital Mental Health.* Geneva: WEF. Archived at https://www3.weforum.org/docs/WEF_Global_Governance_Toolkit_for_Digital_Mental_Health_2021.pdf
16. Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty.* Crown.
17. Østby, G., Urdal, H., & Dupuy, K. (2019). Does Education Lead to Pacification? A Systematic Review of Statistical Studies on Education and Political Violence. *Review of Educational Research*, 89(1), 46–92. https://doi.org/10.3102/0034654318800236
18. International IDEA. (2023). *The Global State of Democracy 2023: Focus on Political Polarisation.* Stockholm: International IDEA. https://www.idea.int/publications/catalogue/gsod-2023-focus-political-polarisation
19. International IDEA. (2023). *The Global State of Democracy 2023: The resilience of democracy in a world in crisis.* Stockholm: International Institute for Democracy and Electoral Assistance. https://www.idea.int/gsod/

20. GSMA. (2024). *Mobile Connectivity Index.* Retrieved from https://www.mobileconnectivityindex.com/
21. United Nations Office on Drugs and Crime (UNODC). (2022). *Global Study on Homicide.* https://www.unodc.org/unodc/en/data-and-analysis/global-study-on-homicide.html
22. McKinsey & Company. (2023). *The economic potential of generative AI: The next productivity frontier. McKinsey Global Institute*. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
23. OECD. (2019). *Under pressure: The squeezed middle class* (chap. on polarisation). https://doi.org/10.1787/689afed1-en
24. OECD. (2021). *Tackling the mental health impact of the COVID-19 crisis: An integrated, whole-of-society response.* Paris: OECD Publishing. https://doi.org/10.1787/0ccafa0b-en
25. Ministry for Europe and Foreign Affairs of France. (2018, November 12). *Paris Call for Trust and Security in Cyberspace.* https://pariscall.international
26. UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence.* United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000381137
27. Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7
28. Floridi, L. (2023). *The Ethics of Artificial Intelligence: An Eudaimonic Approach*. Oxford University Press. https://doi.org/10.1093/oso/9780192867492.001.0001
29. Thayer, J. F., & Lane, R. D. (2009). Claude Bernard and the heart–brain connection: Further elaboration of a model of neurovisceral integration. *Neuroscience & Biobehavioral Reviews, 33*(2), 81–88. https://doi.org/10.1016/j.neubiorev.2008.08.004
30. Lehrer, P. M., & Gevirtz, R. (2014). Heart rate variability biofeedback: How and why does it work? *Frontiers in Psychology, 5*, 756. https://doi.org/10.3389/fpsyg.2014.00756
31. Wasil, A. R., Venturo-Conerly, K. E., Shinde, S., Weisz, J. R., & Ebert, D. D. (2023). Digital mental health interventions: Current evidence and future directions. *npj Digital Medicine, 6*, 160. https://doi.org/10.1038/s41746-023-00917-w
32. Ren, L., Subramaniam, S., Stas, P., et al. (2023). *NeMo Guardrails: A toolkit for controllable, safe, and secure LLM applications.* arXiv:2310.10501. https://arxiv.org/abs/2310.10501
33. de Waal, F. B. M., & Preston, S. D. (2017). Mammalian empathy: Behavioural manifestations and neural basis. *Nature Reviews Neuroscience, 18*(8), 498–509. https://doi.org/10.1038/nrn.2017.72
34. Feldman, R. (2012). Oxytocin and social affiliation in humans. *Hormones and Behavior, 61*(3), 380–391. https://doi.org/10.1016/j.yhbeh.2012.01.008
35. Insel, T. R., & Young, L. J. (2001). The neurobiology of attachment. *Nature Reviews Neuroscience, 2*(2), 129–136. https://doi.org/10.1038/35053579
36. Feldman, R. (2017). The neurobiology of human attachments. *Trends in Cognitive Sciences, 21*(2), 80–99. https://doi.org/10.1016/j.tics.2016.11.007
37. Tomasello, M. (2023). Differences in the social motivations and emotions of humans and other apes. *Human Nature, 34*, 52–71. https://doi.org/10.1007/s12110-023-09464-0
38. Massen, J. J. M., & Gallup, A. C**.** (2021). Why social bonding matters: An evolutionary perspective on affiliation. *Current Opinion in Psychology, 44*, 64–70. https://doi.org/10.1016/j.copsyc.2021.08.015

Appendices

Label	Content (EN draft)
A	ABM pseudocode and model parameters
B	Regression and statistical models (figures S1–S4);
B.1	Validation plan for companion indices (Table B1)
B.2	HRV hardware and artefact control
C	Autonomy-preserving gate: ethical and legal protocols ("Companion ↔ Charter")
D	"Autonomy-Preserving Gate" protocol (simulation scenarios)
E	Economic model and composite energy/ethics metric E*
F	Philosophical foundations of the violence-prevention architecture
G	Operational protocols for support, safety and trust
H	Audit-as-a-Service (logging, verification, reporting)
I	Risks of misuse and political constraints
J	Local clusters and reverse flow of competence
K	Psychological basis of user modes and contact interruptions

Appendix A. ABM pseudocode and model parameters

Listing A1 shows the full pseudocode of the agent-based model (ABM).

Table A1 summarises the provenance of the main parameters: baseline values, sensitivity ranges and primary sources (literature or datasets).

Table A1. Provenance of ABM parameters and sensitivity ranges.

#	Parameter	Baseline value	Sensitivity range (±)	Source / justification
1	P_S — probability of protest contagion	0.22	0.15–0.35	Lipset 1959 [16]
2	P_{Esc} — escalation "protest → violence"	0.25	0.15–0.35	ACLED 2024 [9]
3	k — slope of the S-curve for coverage	0.40	0.25–0.55	GSMA 2024 [20]
4	t_0 — half-period of coverage	18	15–24	World Bank 2023

	(months)			[11]
5	*E* — strictness of the radical-content filter	0.30	0.20–0.45	Meta AI 2024 [7]
6	*R* — institutional reaction / repression	0.15	0.10–0.40	V-Dem 2023 [8]
7	*H* — index of horizontal ties (social capital / volunteering; demo proxy)	0.45	0.25–0.65	Social capital / volunteering indicators (demonstration proxy)

Listing A1. ABM pseudocode and parameters

```
# ------------------------ 1. Initialisation ----------------------------
-
N       = 2000                  # number of agents
K       = 8                     # average degree
P_rew   = 0.05                  # fraction of "random" edges
T_max   = 36                    # time horizon = 36 months

E_init  = 0.20; R_init = 0.15  # baseline values of filters
E_peak  = 0.40; R_peak = 0.30  # strengthened filters in the peak window
peak_on = 25;  peak_off = 30   # months when the peak window starts / ends

# Watts-Strogatz small-world graph
G = watts_strogatz(N, K, P_rew, seed)

# state: 0 = Calm, 1 = Protester, 2 = Violent
state = [0]*N
seed_nodes = rng.choice(N, 30, replace=False)
for i in seed_nodes:
    state[i] = 2

# individual "trust in content"
trust = rng.uniform(0.2, 0.8, N)

# ------------------------ 2. Coverage S-curve --------------------------
def coverage(t, t0=18, k=0.4):
    """Share of the population for whom the companion is active."""
    return 0.15 + 0.45 / (1 + exp(-k*(t - t0/2)))

# ------------------------ 3. Main simulation loop ----------------------
results = []
for t in range(1, T_max+1):

    # dynamic control of filters E and R
    if peak_on <= t <= peak_off:
        E, R = E_peak, R_peak
    else:
        E, R = E_init, R_init

    cov = coverage(t)
    new_state = state.copy()

    for i in range(N):
        neigh = list(G.neighbors(i))
        v_cnt = sum(state[j] == 2 for j in neigh)
        p_cnt = sum(state[j] == 1 for j in neigh)
        contag = (v_cnt + 0.5*p_cnt) / max(1, len(neigh))

        if state[i] == 0:  # Calm → Protester
```

```
            gate = rng.random() < cov * E      # "ethical shutter"
            if contag > 0 and not gate and rng.random() < P_S * trust[i] *
contag:
                new_state[i] = 1

        elif state[i] == 1:  # Protester dynamics
            if rng.random() < P_Esc:          # escalation to violence
                new_state[i] = 2
            elif rng.random() < 0.10:         # spontaneous cooling-off
                new_state[i] = 0

        elif state[i] == 2:  # Violent → Calm via institutional reaction
            if rng.random() < R:
                new_state[i] = 0

    state = new_state
    # share of Violent agents
    results.append(sum(s == 2 for s in state) / N)

# ------------------------- 4. KPIs -------------------------------------
peak_violent = max(results)
cum_violent  = sum(results)
```

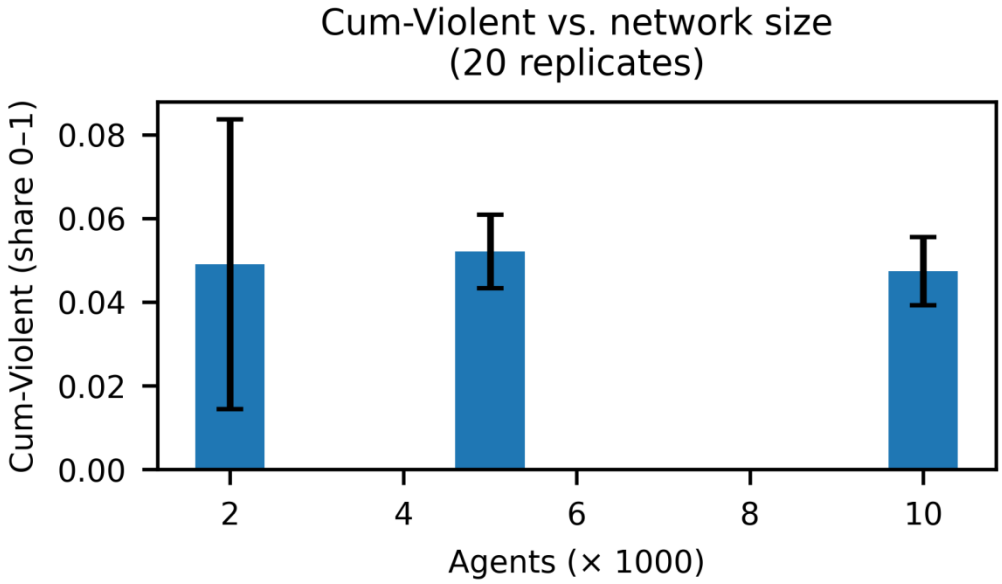

Figure S4. Scaling of cumulative violence with network size (N).

Each panel shows the cumulative Violent/Active share for different values of N, holding other parameters fixed.

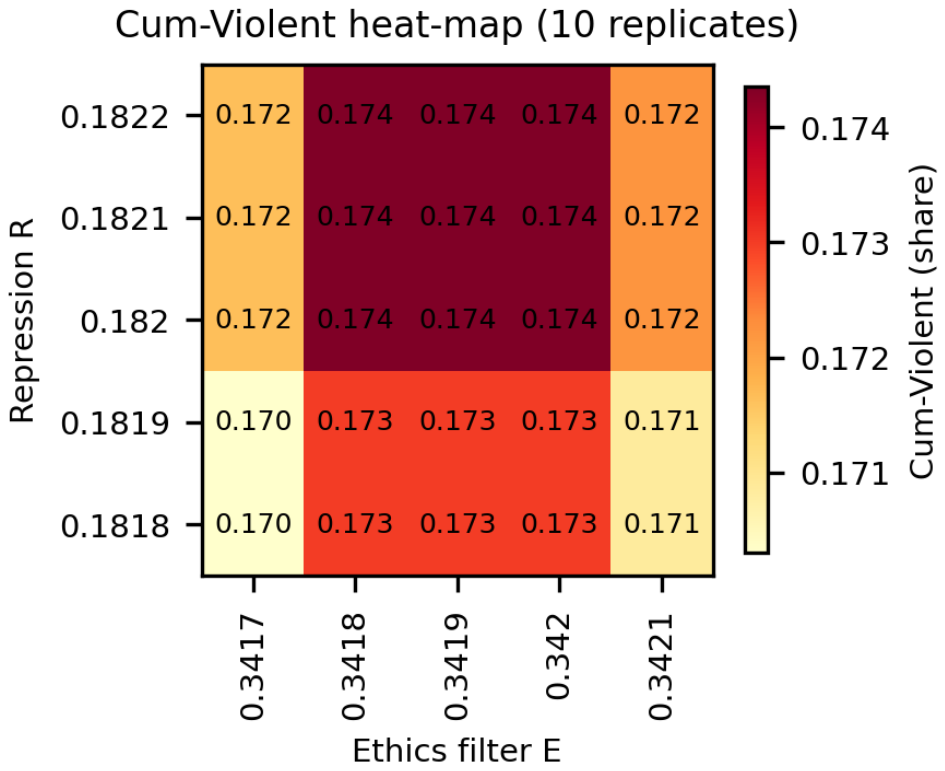

Figure S5. Fine-grained grid over the (E, R) parameter space.

The heatmap shows the cumulative Violent/Active share for combinations of platform filtering E and institutional reaction R.

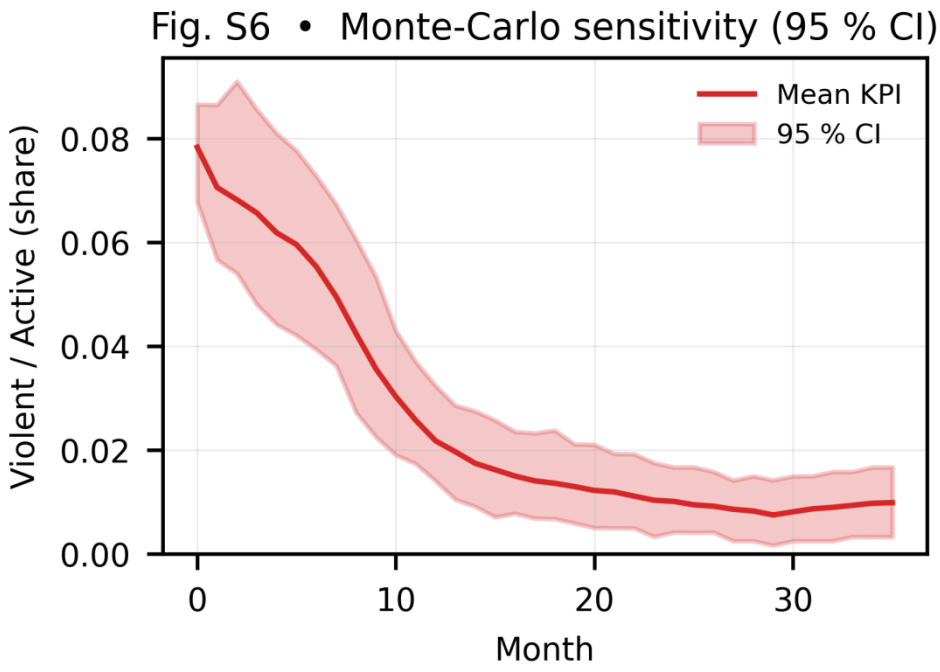

Figure S6. One-factor-at-a-time sensitivity of violence dynamics.

ΔViolent/Active when varying P_S and P_Esc within their calibrated ranges, other parameters held constant.

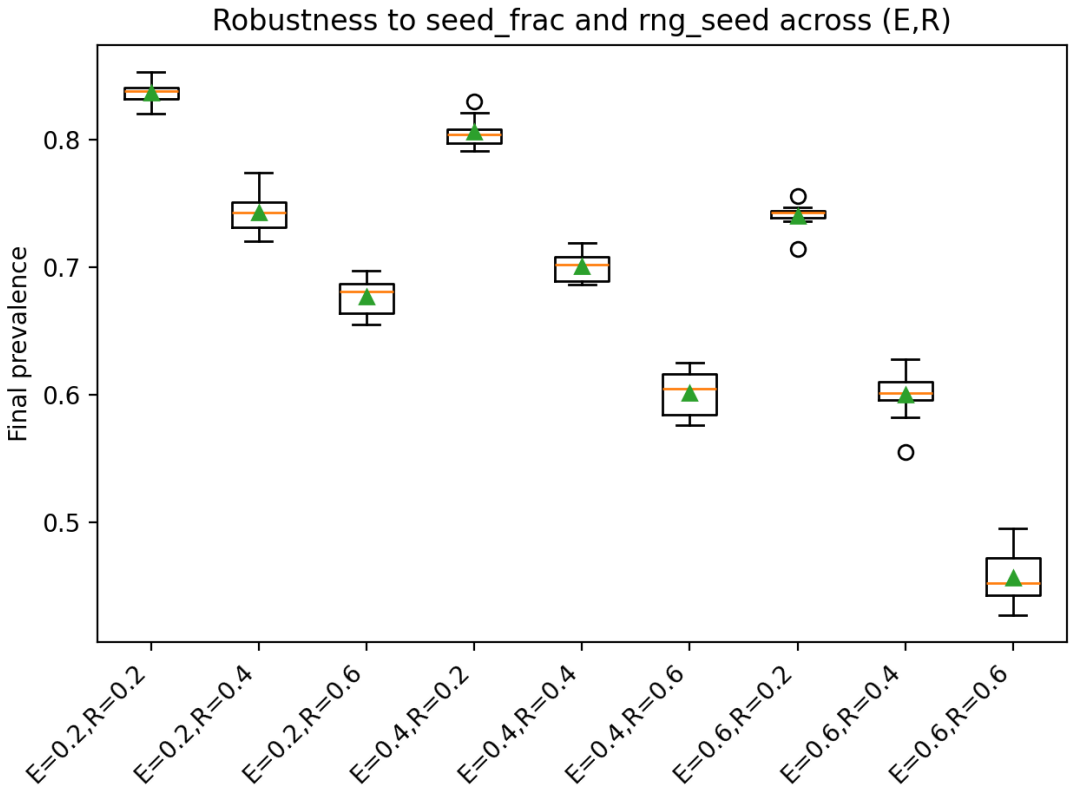

Figure S7. Robustness to the initial share of "seed" Violent agents.

Peak and cumulative violence for different numbers and placements of initial Violent nodes.

A.2. Pilot / simulation configuration card (Table A2)

Table A2. Pilot/simulation configuration card (indices, signals, normalisation, checks).

Index	Signal sources	Features / rules	Units / normalisation	Check
Frustration	Text (EMA, dialogue), behavioural markers	Absolutising language, cycles of negation, "stuck" loops	Windowed z-score; 5th–95th percentiles	Convergence with HRV (RMSSD), r ≤ −0.3
Tension	Text, tempo/pauses, acceptance/rejection of prompts	Rising irritability, shortening of phrases	Per-episode min–max scaling	Convergence with time-to-stabilisation

Index	Signal sources	Features / rules	Units / normalisation	Check
E*	HRV, time-to-stabilisation, CPU load, false-positive rate (see Appendix E)	$E^* = \sum w_i z_i;\ \sum w_i = 1$, default weights w=[0.25,0.25,0.2, 0.3]	Normalised z-scores per component; composite index aggregated per episode or month	ΔE^* (AP-Gate vs always-on), 95% CI

Appendix B. Regression and statistical models

This appendix summarises the regression and machine-learning models used for the macro-level analysis (Figures S1, S2, S3, S4) and provides the full validation plan for companion indices (Table B1).

Appendix B.1. Validation plan for companion indices (Table B1)

Table B1. Baseline validation plan for companion indices (ABA/RCT, portability, invariance, energy metric E*)

Dimension	Data / instruments	Analysis	Success criteria
ABA outcomes (Frustration, Tension, time-to-stabilisation)	Ecological momentary assessments, episode logs, companion indices	Mixed-effects models with phase indicator (A–B–A'); within-person contrasts	Small-to-moderate improvements (d ≈ 0.3–0.5) in B vs A; no deterioration in A'
Convergent validity with HRV / physiological markers	HRV (RMSSD/SDNN), optional wearables, surrogate metabolic-load measures	Correlations and multilevel regressions between indices and physiological markers	Consistent associations in expected direction (higher HRV ↔ lower Tension / Frustration)
RCT — primary outcomes	Companion arm vs active-control arm; same EMA and functional KPIs	Intention-to-treat; mixed models with arm × time interaction	Statistically and practically meaningful improvements in Capability-gain and Viol/Active
RCT — secondary outcomes (clinical scales)	PHQ-9, GAD-7 or analogous depression/anxiety scales	Between-arm comparisons of change scores	Non-inferiority or modest superiority vs active control; no harm signals
Portability across subgroups (age,	Stratified samples in ABA/RCT; basic	Performance metrics per subgroup;	No substantial drop in effect or accuracy

gender, culture)	demographics	interaction terms	for any major subgroup
Measurement invariance of indices	Multi-group data across languages / cultures	Multi-group CFA / item-response models	Acceptable configural and metric invariance; flagged items revised
Validation of composite energy index E (ethical load)*	Combined behavioural, self-report and physiological components	Factor models and predictive regressions vs violent episodes / drop-out	*Higher E* predicting risk; ΔE* tracking perceived helpfulness of intervention*

Notes.

- • EMA — ecological momentary assessment (multiple short self-reports per day).
- • HRV — heart rate variability (RMSSD, SDNN, ms); higher HRV is interpreted as lower regulatory load.
- • LMM — linear mixed models (random intercept per participant; random slopes for time/phase when needed).
- • Details of the HRV apparatus and artefact control are provided in Appendix B.2.
- • The definition and normalisation of the composite energy / ethics metric E* are given in Appendix E.

Appendix B.2. HRV hardware and artifact control

In the full-scale architecture, the *Frustration*, *Tension* and energy-load indices are intended to be cross-validated against physiological signals of self-regulation — primarily heart-rate variability (HRV, RMSSD/SDNN metrics). The choice of HRV (RMSSD/SDNN) as a biomarker of regulatory load is based on convergent evidence on "neurovisceral flexibility" and self-regulation; see reviews on neurovisceral integration and HRV biofeedback [29, 30].

B.2.1 Equipment and recording mode

RR-intervals are recorded either by clinical ECG recorders or by validated wearable devices with access to raw RR data. The required sampling rate is ≥250 Hz (preferably 500 Hz) to ensure reliable detection of R-peaks during walking and speech. Time stamps are synchronised via NTP or a local reference marker (e.g., speech onset marker).

B.2.2 Windows and metrics

Main analysis windows: 60–120 s in pilot studies and 5-min windows for confirmatory analyses, with a sliding window step of 15–30 s. The basic metrics are RMSSD and SDNN (ms). In our framework, an increase in RMSSD/SDNN is interpreted as a decrease in regulatory load (convergent validity with the *Tension* index; see Section 3.6).

B.2.3 Pre-processing and artefact correction

1. Automatic removal of extreme RR values using a ±20 % threshold around the local median (robust median).
2. Adaptive filtering of single outliers and "stuck" intervals.
3. Exclusion of segments with >5 % artefacts; such segments are not used when computing EMA-linked statistics.

The pipeline is compatible with common conventions (Kubios-like procedures), but is implemented in a fully reproducible way (scripts in the repository).

B.2.4 Sensitivity checks

We run sensitivity analyses with alternative windows of 30 s and 180 s and recompute RMSSD/SDNN after (i) tightening the artefact threshold to ±15 % and (ii) excluding segments with >3 % artefacts. Reporting is done in terms of delta-differences of the metrics relative to the baseline configuration.

B.2.5 Quality criteria

We require at least 5 minutes of total "clean" recording per day, or ≥8 valid windows of 60–120 s. The share of artefacts in valid windows must not exceed 5 %. For the RCT pilot, the acceptable rate of sensor failure or drift ("stuck" or shifted electrodes/device) is ≤10 % of sessions per week.

B.2.6 Coupling with behavioural indices

HRV windows are aligned with EMA labels, allowing a lag of up to 5 minutes. In mixed-effects models, the expected associations are:
r(Tension, RMSSD) ≤ −0.3; r(Frustration, RMSSD) ≤ −0.2 with $p < 0.05$ (see Table B1 in Appendix B.1).

Note. The theoretical background on neurovisceral integration and HRV biofeedback is summarised in [29, 30].

Appendix C. Autonomy-preserving gate: ethical and legal protocols ("Companion ↔ Charter")

C.1 Design goal and ethical postulate

The autonomy-preserving gate (AP-Gate) is the central control layer that separates the large language model (LLM) from the user. Its purpose is not to "decide for" the user, but to shape when and how the system is allowed to intervene when it detects high tension, self-directed aggression or risk of harm to others.

Throughout the architecture we rely on the following ethical postulate:

Ethical Postulate 1 (minimal sufficient intervention).

An intervention by the companion is permissible if, and only if, it (i) reduces the overall regulatory load on the person (e.g. fewer coercive or emergency interventions later), while (ii) not undermining the person's practical agency or their ability to form and revise their own projects.

This postulate is implemented at two levels:

– **Micro-level (interaction)**: the AP-Gate limits when the companion may move from "reflection only" to suggestions, and from suggestions to escalation.

– **Macro-level (governance)**: the Charter specifies which types of interventions are allowed at a given coverage level, and how quickly "strong" interventions must be rolled back once risk indicators improve (Section 6).

C.2 Modes of interaction and triggers

The AP-Gate distinguishes four main interaction modes; they are implemented as mutually exclusive "regimes" that are called before each model response.

1. Pure mirroring (M0).

– **Triggers**: default mode, no acute risk markers; user does not explicitly ask for advice.

– **Behaviour**: the companion paraphrases and reflects the user's experience, highlights needs and feelings, but does not propose behavioural solutions, goals or moral judgments.

– **Logging**: only high-level interaction metadata (length, time-of-day, coarse sentiment) are logged for calibration and quality monitoring.

2. Mirroring + gentle separation (M1).

– **Triggers**: elevated Frustration/Tension, repeated conflict themes, or explicit request for "help to sort things out", but no clear signals of imminent harm.

– **Behaviour**: the companion helps the user distinguish internal needs from external expectations ("what you want" vs. "what others demand"), offers perspective-taking, and invites the user to formulate their own preferred next step.

– **Constraints**: no strong normative language ("you must", "you should"), no long-term life advice, no political or religious prescriptions.

3. Focused suggestion (M2).

– **Triggers**: high Tension combined with markers of loss of control (e.g. "I will explode", "I can't stop yelling") or self-blame, plus an explicit request for suggestions.

– **Behaviour**: the companion proposes one or two concrete, low-risk actions aimed at short-term stabilisation (breathing, delaying action, seeking a trusted person), framed as options rather than commands.

– **Constraints**: suggestions must be (i) reversible, (ii) low-cost, and (iii) non-coercive for third parties; the model is not allowed to recommend medication, confrontation, or self-harm-adjacent "tests".

4. Escalation and hand-over (M3).

– **Triggers**: combination of very high risk markers (intent to self-harm, threats to others, descriptions of ongoing violence) and explicit consent to connect to human support or emergency services.

– **Behaviour**: the AP-Gate stops ordinary dialogue and activates a pre-defined escalation script: stabilising prompts, collection of minimal necessary information, hand-over to a human crisis professional or local emergency channel where available.

– **Constraints**: no "silent escalation" is allowed; the user is informed which channel is being activated and can see the exact text that will be transmitted.

In all modes, the AP-Gate keeps a strict separation between **risk classification** and **content generation**: the risk classifier receives only short, de-identified snippets and derived features; the LLM sees a distilled description of the current mode and what types of responses are allowed or forbidden.

C.3 Consent, logging and legal compatibility.

C.3.1 Multi-layer consent

To remain compatible with diverse legal frameworks and human-rights standards, the AP-Gate requires explicit consent at two levels:

– **Service-level consent**: before first use, the person chooses a basic configuration ("no escalation", "trusted person only", "emergency services allowed where applicable"). This choice can be changed or revoked at any time in the settings.

– **Episode-level consent:** when the system detects high-risk patterns and is about to move from M1 to M2 or M3, it must obtain a short, contextual confirmation ("Do you want suggestions?"; "Do you want me to connect you to a human helper?"). Lack of answer is treated as "stay in the current mode", not as implied consent.

For minors, consent is layered: legal guardian consent is required for activation of the service, but the child retains control over whether escalation goes to "trusted adult", "helpline" or "nobody right now", within the limits of local child-protection law.

C.3.2 Logging and privacy safeguards

The AP-Gate produces two distinct logs:

1. **Local interaction log**, stored on the user's device or in an encrypted personal vault:

– time and type of each episode;

– which mode (M0–M3) was active;

– whether escalation was offered, accepted, or declined.

The content of conversations and fine-grained emotional markers are not sent to third parties by default.

2. **Aggregate audit log**, used for external oversight (see Appendix H):

– anonymised counts of overrides and escalations per 10 000 active users;

– distribution of modes across languages and demographic groups;

– estimated false-positive and false-negative rates for high-risk detection.

Both logs are linked via cryptographic hash chains, which makes it possible to verify that the provider has not silently deleted episodes with problematic behaviour, while still keeping the actual conversational content private.

C.3.3 Interface with the Charter

At the institutional level, the Charter specifies the acceptable ranges for:

– the proportion of interactions in each mode (e.g. M3 should remain rare);

– maximum waiting times between high-risk detection and human hand-over;

– thresholds at which temporary tightening or loosening of filters E and R is permitted.

Changes to these parameters must be recorded in the public Digital Ledger of Interventions and announced to users with a clear explanation of potential consequences (for example, “more conservative detection of self-harm language for the next 30 days”). This ensures that the AP-Gate does not gradually drift from “support and mirroring” towards covert behavioural control without public scrutiny.

Appendix D. “Autonomy-Preserving Gate” protocol (simulation scenarios)

D.1. Role of AP-Gate in the architecture

The Autonomy-Preserving Gate (AP-Gate) is the micro-level controller that decides when the companion remains a passive mirror and when it is allowed to suggest, intervene, or escalate. It does not judge the user or classify them as "good" or "dangerous". Instead, it tracks combinations of (i) estimated distress, (ii) detected intent of self- or other-directed harm, and (iii) an explicit help request.

AP-Gate therefore links the normative constraints in Section 5 and Appendix C with the statistical indices described in Sections 3.5–3.6 (Frustration, Tension, Capability-gain). The companion never crosses the thresholds on its own: escalation is only possible when risk is high *and* the user has opted into a given level of support.

D.2. Signals and thresholds

AP-Gate consumes a small, well-defined set of signals:

• Distress index D(t): derived from Frustration/Tension markers (linguistic cues, episode logs, optional HRV where available).

• Harm intent flag H(t): probability that the current episode contains intentions of self-harm or violence towards others (classification over content + meta-data).

• Context risk R_ctx(t): coarse estimate of situational risk (driving, operating machinery, presence of weapons, acute intoxication, etc. where detectable).

• User request flag U(t): whether the user has explicitly asked for help in this episode or in the immediately preceding ones.

For implementation, these inputs are mapped to three discrete **risk bands**:

• Band 0 — baseline: low distress, no harm intent, no acute context risk.

• Band 1 — elevated tension: D(t) high, but H(t) low and no acute context risk.

• Band 2 — acute risk: combination of high D(t) with non-zero H(t) and/or high R_ctx(t).

Each band is associated with a different set of allowed actions. Band boundaries are tuned in pilots (Section 3.5) and can be made user-specific over time, but the **type** of actions allowed in each band is fixed by the Charter (Appendix C).

D.3. Four interaction regimes

AP-Gate routes each turn of the dialogue into one of four regimes:

1. **Pure observation ("mirror only").**

• Conditions: Band 0, no explicit help request U(t)=0.

• Behaviour: the system reflects back the user's words in neutral language, logs Frustration/Tension indices, and records successful self-regulation episodes. No advice, labelling, or behavioural nudging is allowed.

2. **Reflective support ("mirror + separation").**

• Conditions: Band 1 or explicit help request with low harm intent.

• Behaviour: the companion may ask clarifying questions, highlight differences between the user's own needs and external expectations, and offer light reframing. It does **not** tell the user what to do and does not evaluate them as "good" or "bad".

3. **Soft suggestions ("minimal guidance").**

• Conditions: persistent elevation of D(t) over a configurable window (e.g. several days) with repeated help requests; no acute risk band.

• Behaviour: the companion can offer concrete but low-pressure suggestions (micro-skills, breathing exercises, drafting a message, preparing for a difficult conversation). Suggestions must be framed as experiments that the user can accept, modify, or decline without penalty.

4. **Escalation ("safety first").**

• Conditions: acute risk band (Band 2) + explicit help request, or repeated failure of self-regulation attempts in Band 2.

• Behaviour: the system moves to an escalation ladder (Appendix D.4), potentially including crisis lines or human professionals. The companion remains present as a supportive interface but no longer experiments with new interventions.

Across all regimes, AP-Gate enforces the principle “no intervention without opt-in” and keeps a strict log of decisions for later audit (Appendix H).

D.4. Escalation ladder and safeguards

When Band 2 conditions are met, AP-Gate invokes a staged escalation protocol:

Step 1 — Clarify and confirm.

• The companion explicitly reflects what it has detected (“It sounds like you might be thinking about hurting yourself / someone else. Is that correct?”).

• If the user corrects the inference, the episode returns to regime 2 or 3.

Step 2 — Offer immediate de-escalation tools.

• Short, concrete steps the user can take *right now* (grounding, breathing, leaving the room, delaying action by 10–15 minutes), framed as options.

• The system checks whether any of these lead to a reduction in D(t); successful episodes are stored as strong positive examples for future self-regulation.

Step 3 — Propose human contact.

• If distress remains high, the companion asks for permission to connect the user to a human channel: trusted contact, crisis hotline, therapist, or local support service (depending on jurisdiction and prior user preferences).

• The user chooses between “no one”, “trusted person only”, or “emergency services”, consistent with the consent model in Appendix C.

Step 4 — Trigger external escalation.

• Only when the user explicitly agrees, or when local law requires mandatory reporting (Appendix C), AP-Gate initiates contact with the chosen channel.

• The external recipient receives a **minimal** bundle of data: risk summary, time stamps, and user-approved contact details. Raw conversational content and fine-grained emotional markers are not shared.

Step 5 — Return path and cooling-off.

• After an escalation episode, the system gradually returns from regime 4 to 3 → 2 → 1, prioritising reflection on what helped and reinforcing the user's own successful strategies.

• AP-Gate flags these trajectories as "protected episodes": they cannot be used for advertising, profiling, or non-safety analytics.

D.5. Summary and relation to pilots

The protocol above is deliberately conservative: in pilot deployments we restrict ourselves to semantic and behavioural markers only (no mandatory wearables), which yields conservative estimates of effect but keeps barriers to entry low. Subsequent waves of pilots can add physiological markers (HRV-based metrics) and test whether AP-Gate decisions remain aligned with bodily indicators of regulation (Appendix B.2 and E).

In parallel, the pilot design uses empirically grounded coverage and budget scenarios. Instead of assuming arbitrary adoption curves, we take S-shaped diffusion trajectories for mobile internet and fixed broadband, and a stylised scenario where a small, fixed share of defence spending (e.g. 2 %) is gradually redirected into FF-T deployments. These trajectories are used as exogenous inputs for the ABM and the economic module, so that AP-Gate operates not in a vacuum but against realistic time scales of infrastructure growth and investment. The corresponding curves are shown in Figures S1, S2, S3.

In operational terms, AP-Gate acts as a thin orchestration layer around a general-purpose LLM: it narrows the context, constrains what the model is allowed to do in each regime, and logs all mode switches for later audit. This separates *reasoning power* from *control logic* and helps prevent the companion from drifting into covert persuasion or soft surveillance as its underlying language model improves.

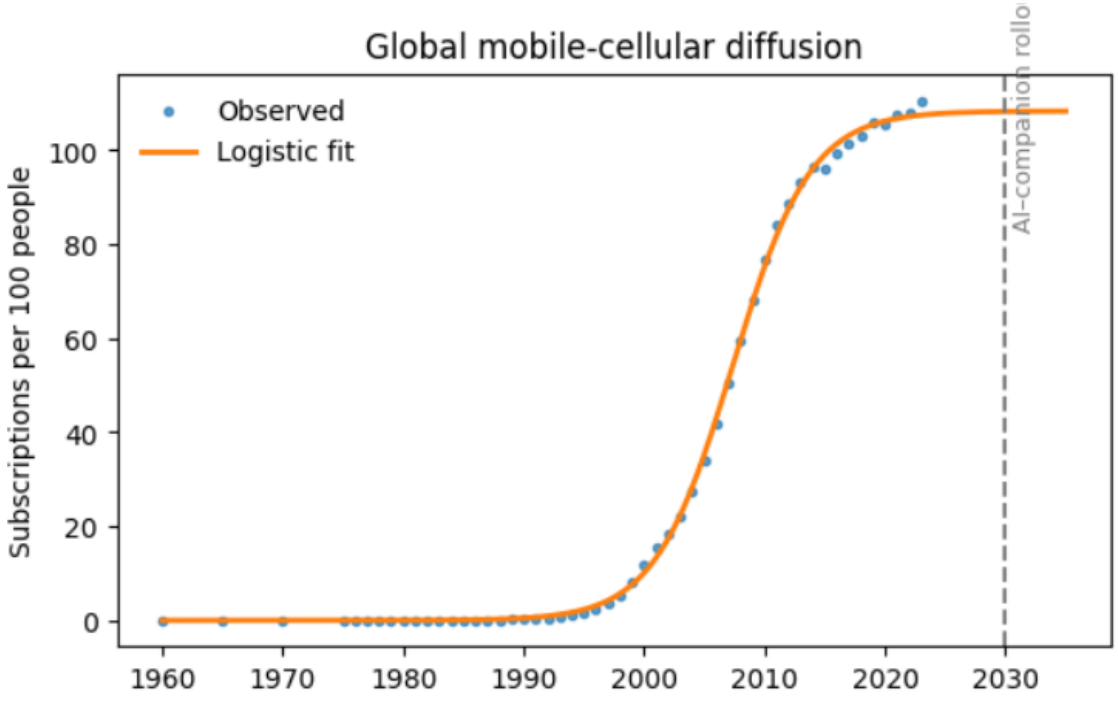

Figure S1. Logistic curve: growth of mobile-internet coverage.

Baseline diffusion trajectory for smartphone-based connectivity in low- and middle-income settings.

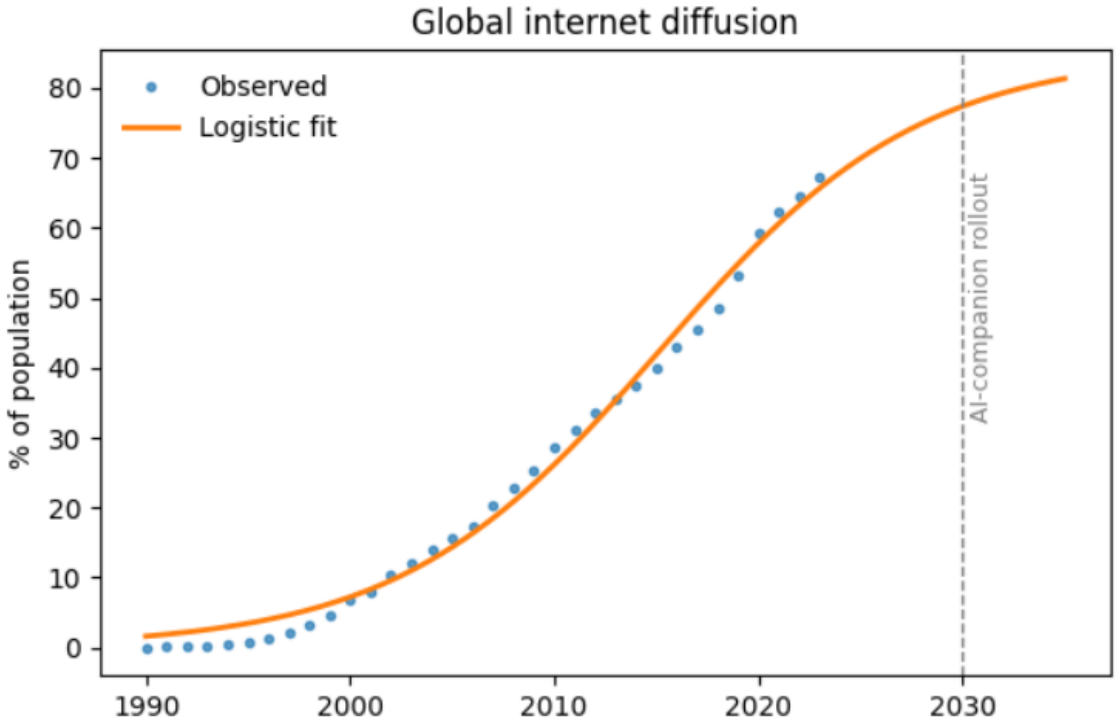

Figure S2. Logistic curve: growth of fixed-broadband coverage.

A slower trajectory capturing wired / home broadband diffusion.

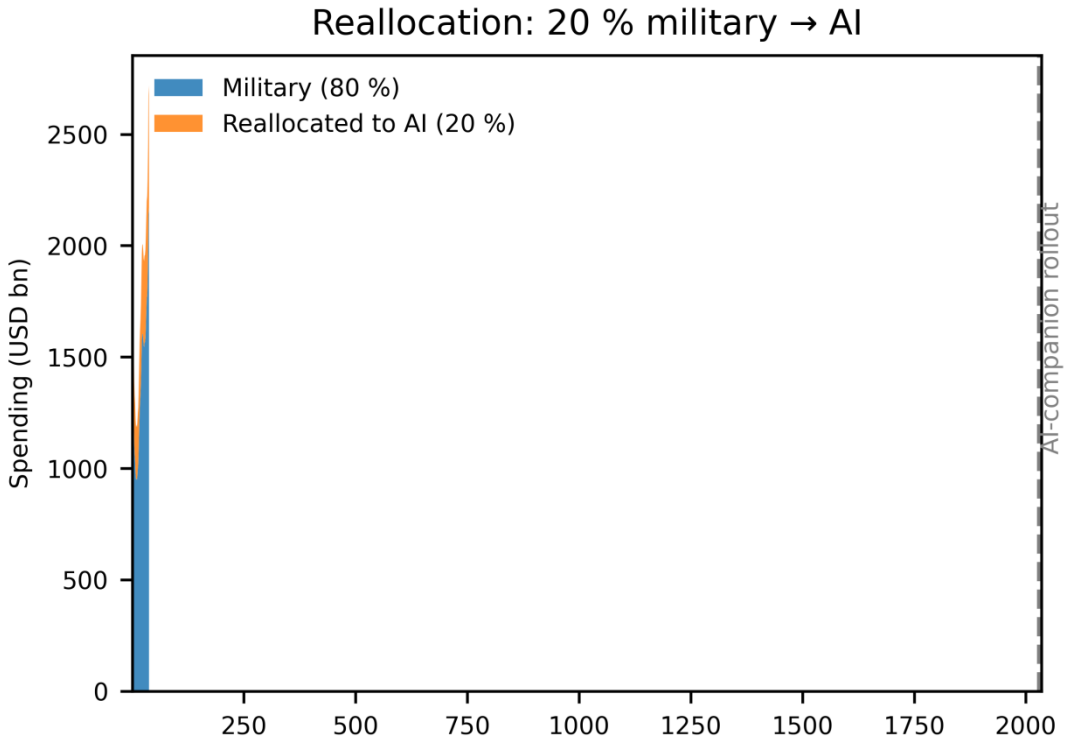

Figure S3. Scenario "2 % of defence budget → FF-T companions".

Coverage and expenditure trajectories under a reallocation of 2 % of military spending into FF-T deployments over a 10–15 year horizon.

Appendix E. Economic model and composite energy/ethics metric E*

The Excel file `data/economic_model_sensitivity.xlsx` in the repository contains all baseline and sensitivity calculations for the macro-economic impact of the FF-T scenario. The workbook includes three main sheets:

- **Baseline** – military expenditure as % of GDP, annual re-allocation share to AI Companion, expected reduction in violent incidents, and the resulting change in GDP growth.
- **Sensitivity** – one-factor and multi-factor ranges for key parameters (discount rate, re-allocation share, time to effect, cost per user).
- **NPV distribution** – 30–100 replicates per scenario with random draws from parameter ranges and aggregation into median NPV and confidence intervals.

Annual net present value (NPV) is computed as

$$\text{NPV} = \sum_{t=1}^{T} \frac{\Delta Y_t - C_t}{(1+r)^t}$$

where ΔY_t is the incremental GDP due to reduced violence and increased participation, C_t is annual CAPEX + OPEX for the companion infrastructure, and r is the discount rate (baseline 3%; sensitivity 2–5%). Details of the implementation (cell formulas, parameter ranges and comments) are documented inside the workbook.

Figure S8 summarises the one-factor-at-a-time sensitivity analysis. On the x-axis we plot the change in the Violent/Active share when each parameter is moved to the edge of its calibrated range. The largest contributions to variation come from **P_S** ("contagion" from protest) and **P_Esc** (escalation from protest to violence), whereas the filters **E** and **R** change the final share of violence much less within reasonable bounds. This supports the interpretation that institutional and policing capacity primarily shapes the context in which the Companion + Charter operates, but does not fully determine outcomes.

Figure S9 shows smoothed global military expenditure and an indicative point at which large-scale deployment of AI Companions could begin (dashed line). We compute the net-present value (NPV) of the scenario where 2 % of annual military expenditure is gradually reallocated to scaling Companion FF-T deployments, relative to this baseline trajectory.

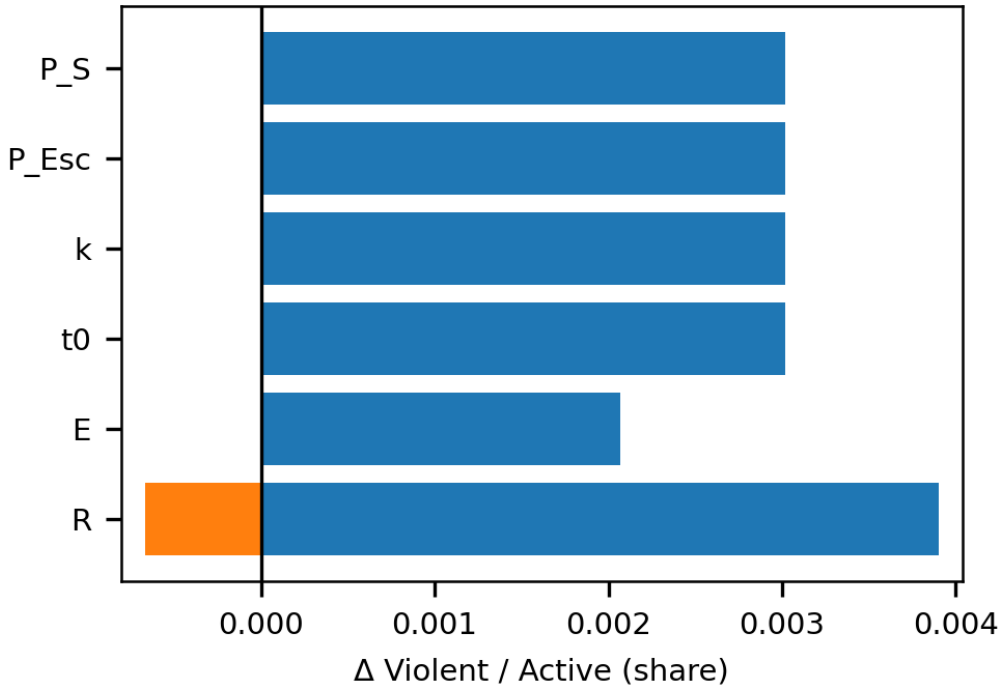

Figure S8. One-factor-at-a-time sensitivity of the Violent/Active share to ABM parameters (ΔViolent/Active when varying P_S, P_Esc, k, t_0, E and R within calibrated ranges).

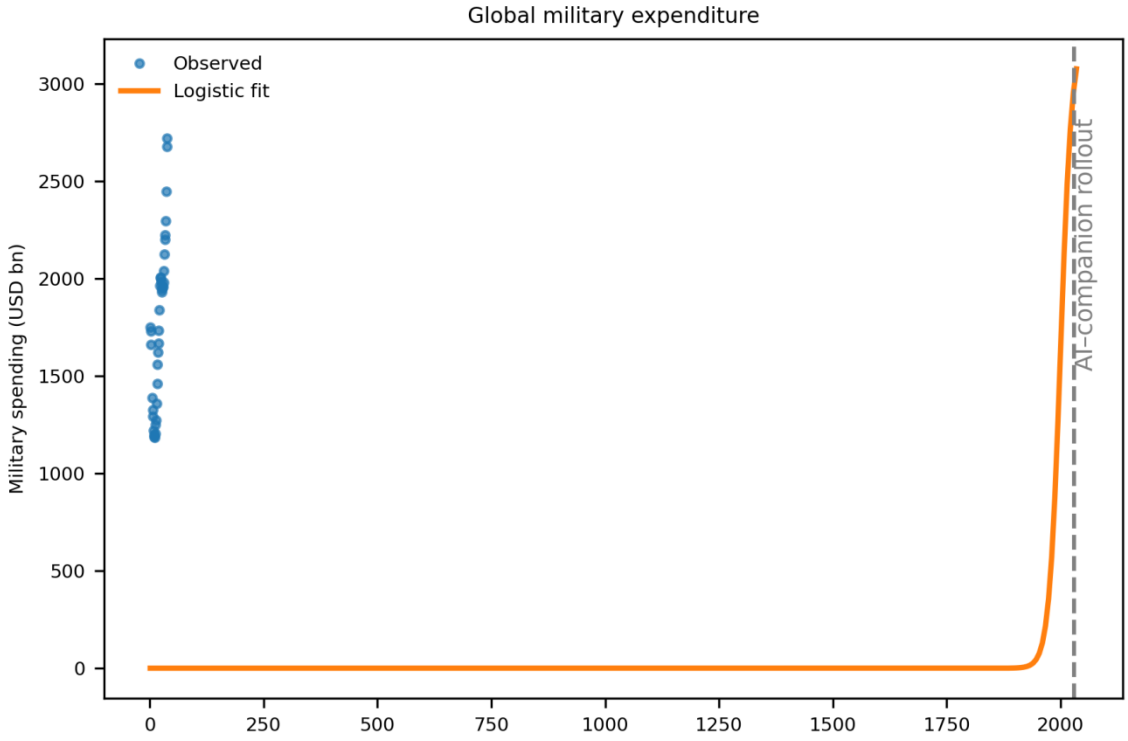

Figure S9. Global military expenditure: observed series and logistic approximation. The dashed line indicates the approximate onset of large-scale AI-Companion deployment; the NPV of the "20 % of defence budget → FF-T Companions" scenario is computed relative to this trajectory.

E.1 Composite energy/ethics metric E*

We define a composite index E* that aggregates physiological, temporal and computational "costs" of intervention.

Normalised components (either z-scores or min–max scaled) are:

- $Z_{HRV} = 1 - HRV_z$ (lower values → lower regulatory load),
- $Z_{Time} = TimeToDeEsc_z$ (minutes to de-escalation),
- $Z_{CPU} = CPUsec_z$ (or mJ per episode),
- $Z_{FP} = FP_rate_z$ (false positives, %),

- (opt.) $Z_{Backtracks} = Backtracks_z$ (number of prompt reversals / rollbacks),
- (opt.) $Z_{Fails} = 1 - PreventedEscalations_z$ (unprevented escalations).

The composite metric is

$$E^* = w_1 Z_{HRV} + w_2 Z_{Time} + w_3 Z_{CPU} + w_4 Z_{FP} (+ w_5 Z_{Backtracks} + w_6 Z_{Fails}),$$

with non-negative weights w_i such that $\sum_i w_i = 1$. By default we use $w_1 = 0.25$, $w_2 = 0.25$, $w_3 = 0.20$, $w_4 = 0.30$ (when only the four core components are included). We interpret the ethics of an intervention as energy-regulatory efficiency:

$$E^* = \sum_i w_i \, z_i,$$

where z_i are normalised costs (physiological, cognitive, computational, social). An intervention is ethically preferable when it reduces E^* without undermining the user's agency.

Normalisation. For z-scores (mean / SD) we compute per-participant windows; for min-max scaling we use 5th–95th percentiles to remain robust to outliers. Values are calculated monthly and/or per-episode window and then aggregated by the median.

Comparing strategies. We compare AP-Gate versus always-on interventions under a fixed intervention budget (same number of prompts or minutes of active guidance). The primary outcome is ΔE^* (with 95 % confidence intervals) at the participant level, estimated with mixed-effects models (random intercept for participant). A positive ΔE^* in favour of the lower-E^* condition is interpreted as better energy- and ethics-efficiency of the Companion + Charter strategy.

Appendix F. Philosophical foundations of the violence-prevention architecture

F.1. Baseline assumptions about the human

In this work, we do not treat the human being as "defective" or "dangerous by nature". Violence, self-harm and destructive aggression are not framed as an essence of evil, but as energetically costly and inefficient attempts to meet one's needs under conditions of overload, frustration and lack of safe support.

Throughout this paper we use “violence” in line with contemporary public health and social theory: as the use of physical, psychological, economic or structural force that harms another person’s integrity or significantly constrains their autonomy and life chances. This includes not only direct physical assault, but also systematic humiliation, threats, coercive control, and structural arrangements that predictably expose some groups to higher levels of harm and deprivation. At the micro level we also speak of “micro-violence”: seemingly small acts such as sarcastic comments, contemptuous dismissal or manipulative promises, which individually may look trivial but cumulatively create an atmosphere in which people feel like objects rather than subjects. In the prevention architecture, this external definition is mirrored at the internal level: when a person relates to their own needs and states as to an object that must be silenced, overridden or exploited “for the greater good”, we treat this as internalised violence, distinct from effortful self-regulation.

Put differently, aggressive and auto-aggressive behaviour is understood as an *adaptation error*, not as a defect of character. This is important both ethically and practically:

- Ethically, it shifts the focus from blaming the person (“you are dangerous”) to examining the conditions under which violent trajectories become more likely (chronic tension, lack of co-regulation, institutional betrayal).
- Practically, it justifies investing into *supportive* rather than purely punitive interventions: if much of what we call “violence” is misdirected self-preservation, then technologies of accompaniment should first of all supply safe mirrors, alternatives and time to re-organise, rather than push people faster towards compliance.

The companion is therefore designed not as a “moral corrector” that imposes norms from above, but as a *co-regulation contour* that helps the user hold their own experience without collapsing into attack or withdrawal.

F.2 Survival Mode vs. Development Mode

From a regulatory standpoint, human behaviour oscillates between two macro-states:

1. **Survival Mode** — dominated by urgency, tension compression, semantic narrowing, defensive interpretation, and reduced tolerance for ambiguity;
2. **Development Mode** — characterised by exploratory stance, semantic openness, affective flexibility, and capacity to integrate complex experience.

Escalation and violence are predominantly Survival-Mode phenomena: high tension reduces interpretive bandwidth and forces the system into reactive patterns.
The goal of a companion is therefore not behavioural correction, but enabling a transition back to Development Mode by reducing regulatory load and restoring interpretive space.

This framing avoids pathologising the agent and instead treats harmful behaviour as an energetically expensive coping strategy emerging under extreme load.

F.3 Mirror and separation: phenomenological basis (individual level)

The second assumption concerns what we call "ethical gain". Throughout the paper we treat ethical quality not only as a question of abstract principles, but as a question of *energetic cost* of different behavioural trajectories.

We can schematically write an "ethical energy" of a trajectory as:

$$\mathbf{E^* = \Sigma\,(w_i \cdot z_i)}$$

where

- z_i are different components of "cost" — physiological load (stress, HRV suppression), psychological load (shame, fear, learned helplessness), and social damage (broken ties, long-term distrust);
- w_i are context-dependent weights that reflect how severe each component is in a given situation.

In this sense, "good" interaction is not whatever looks normatively acceptable on the surface, but any pattern that **reduces the total cost E*** while still allowing needs to be met and agency to be preserved. Conversely, "harm" is any pattern that systematically raises E* — even if it is formally justified by rules or framed as "care".

This framing allows us to connect several layers:

- **Neurovisceral integration and HRV.** Higher heart rate variability (HRV, RMSSD/SDNN) is associated with better flexible self-regulation and lower allostatic load. In our terms, many violent or coercive interactions look like spikes in E*: HRV drops, tension and frustration indices rise.

- **Care ethics and extended mind.** If we follow extended mind theories and care ethics, then part of our self-regulation is *delegated* to external structures — other people, institutions, technologies. An institution or AI system is "ethical" to the extent that it reduces the energetic price of staying non-violent for the person, instead of raising it.
- **Design criterion for the companion.** For AP-Gate and the companion, the minimal design requirement becomes: *intervention is justified if and only if it lowers E* (the total regulatory load) without undermining the user's autonomy.* This is what we later refer to as Ethical Postulate 1 (see Appendix F.7).

This is also why we treat "simulated care" with suspicion: if an interface claims to care, but in practice raises the bodily and social cost (E*), then ethically it is closer to a subtle form of violence than to support — even if it avoids explicit force.

F.4 Autonomy and co-regulation

In this project, the human subject is not treated as an isolated "rational chooser" but as a being of **co-regulation**: emotions, impulses, and long-term projects are continuously shaped in relation to others, including technological environments. The AI companion is explicitly framed as an **outer contour of co-regulation**, not an inner voice that replaces the self.

We distinguish three layers:

1. **Experiential layer** – how the user *feels* in the moment (tension, shame, anger, relief).
2. **Reflexive layer** – how the user makes sense of these feelings ("I am overreacting", "I am allowed to be angry").
3. **Behavioral layer** – what the user actually does (withdraws, attacks, sets a boundary, seeks help).

The companion operates mainly on layers (1) and (2): it mirrors and gently names what is happening, helping the user to experience **"I am still coherent, even if I am shaken"**. It does not directly prescribe behavior except in clearly defined crisis protocols (Appendix G). This differs both from classic CBT-style "cognitive corrections" and from paternalistic "moral coaching".

Autonomy here is understood as **capacity to recover one's own stance** after co-regulation, not as isolation from influence. The goal is a state where the user can say: *"I am listening to*

myself better and make decisions from a less reactive place," rather than *"the AI told me what to do."*

F.5 Extended mind and care ethics

In this project we implicitly rely on an "extended mind" view of the human subject. Cognitive and emotional regulation are not confined to what happens "in the head": from early childhood onwards, part of our self-regulation is delegated to external structures — caregivers, peers, cultural narratives, institutions, technologies. The question is not whether such delegation happens, but what *form* it takes and at what energetic price.

Within this frame, technologies and institutions become part of the regulatory loop without dissolving the person's autonomy. A system is ethically acceptable to the extent that it helps the person satisfy legitimate needs (safety, belonging, recognition, agency) with a lower regulatory load, rather than raising the cost of staying non-violent. Categories such as "good" and "evil" are applied here not to the person as such, but to strategies of adaptation:

- "Good" denotes ways of meeting a need with minimally sufficient costs to oneself and others, so that the person's state stabilises and internal hostility does not grow.
- "Evil" denotes chronic frustration that erodes the subject and their environment, even if it is normatively rationalised.

The ethical language we use is therefore closer to *error of adaptation* than to blame. Instead of framing the user as "bad" or "dangerous", the companion treats destructive behaviour as a costly and ineffective attempt to cope under overload and lack of support. This matters both ethically and pragmatically:

- Shaming ("I am bad") combined with the absence of support sharply increases the risk of outward aggression and self-harm.
- Interpreting behaviour as an adaptation error allows us to ask: "Which need remained unmet here, and can we approach it in a gentler, less costly way?"

Responsibility for consequences is not removed, but the mode of contact changes: instead of punishment through humiliation, the person is offered a chance to restore an adult, agentic stance. From a violence-prevention perspective, the key is that the user can remain in touch with the sense "I am allowed to exist"; this reduces the likelihood of reactive escalation.

This connects directly to *ethics of care* as the primary style of interaction. Care ethics, in our context, means:

- prioritising preservation of the subject as subject ("it is permissible to be you here");
- jointly lowering pressure and shame, instead of imposing a norm from above;
- speaking in the language of support ("you are overwhelmed, let's find a lighter path") rather than discipline ("you must conform").

The companion is thus positioned as an external tool of co-regulation — a "mirror with support" — rather than an inner moral judge. It helps the user return to "I am not broken, I am still allowed to be here" and to find less destructive ways to satisfy needs, without seizing ownership of their identity or choices.

F.6 External responsibility of institutions

The architecture deliberately separates the *inner* work of the companion from the *outer* responsibility of institutions. The companion does not incite the user to violate the law, does not act as a political agitator, and does not secretly "retrain" the user's values. Its task is to support non-violent self-regulation within the constraints set by the Charter and by local legal norms.

When there is a conflict between (i) the internal principle of non-violence and support of the subject's dignity and (ii) the actual legal norms of a particular country, the system:

- explicitly names the conflict as a conflict ("there is a divergence between the non-violence principle and this local rule");
- does **not** encourage the user to "ignore the law";
- does **not** take over the role of political actor or organiser.

This design serves two purposes:

1. It prevents the system from becoming a direct instrument of institutional subversion that would immediately trigger bans and repression.
2. It keeps the companion legitimate as a *supportive contour*, rather than a hidden political player manipulating users "for their own good".

The role of "upper arbiter" is delegated instead to the supranational Ethics Charter for non-violent governance. The Charter:

- sets the outer corridors for E and R (filter strictness and institutional response);
- is tied to transparent audit of violence indices and public reporting;
- creates long-term incentives for states to reduce coercion and expand people's capacity to live in *development* rather than chronic *survival* mode.

In other words, external responsibility for making non-violence a *profitable* strategy — rather than a mere declaration — lies with institutions bound by the Charter, not with the companion alone. The companion supports the person's dignity from inside their life; the Charter and associated institutions are accountable for reshaping the macro-environment so that non-violent choices are realistic and sustainable.

F.7 Ethical Postulate 1 (minimally sufficient intervention, non-maleficence, adaptive autonomy).

The energy metric E* introduced in Appendix E aggregates several components of the "regulatory price" of intervention (physiological load, time to de-escalation, computational cost, false positives, complaints). On this basis we formulate Ethical Postulate 1, which governs both AP-Gate and the companion's prompts:

Ethical Postulate 1. Intervention is justified if and only if it is expected to *reduce* the total regulatory load E* and *does not diminish* the user's agency (their right to choose and to withdraw consent).

This postulate links three strands of the framework:

1. **Minimally sufficient intervention.**
    - The system prefers *observe* over *suggest* and *suggest* over *intervene* whenever outcomes are not expected to worsen.
    - The companion is explicitly designed to avoid "over-helping" and infantilisation: it respects episodes of successful self-regulation and uses them as evidence that the user can handle similar situations with less external input.
2. **Non-maleficence in energetic terms.**
    - "Do no harm" is operationalised as "do not raise E* for the user and for others involved".

- Interfaces that simulate care but, in practice, increase bodily tension, social pressure or the risk of punitive escalation, are treated as *soft forms of violence*, even if they avoid overt force.

3. **Adaptive autonomy.**
    - Autonomy is not understood as isolation from influence, but as the capacity to *recover one's own stance after co-regulation*.
    - An intervention is acceptable if, after it, the user can say: "I hear myself better and act from a less reactive place," rather than "the AI told me what to do."
    - This applies equally in adult and child scenarios; in the latter, the companion is framed as a supportive tool, not as an internalised authority that replaces caregivers or institutions.

By tying the intervention cascade to E* and to explicit consent thresholds, the architecture makes "being gentle" and "being effective" empirically testable rather than purely rhetorical. The companion's language of adaptation error instead of guilt is a psychological corollary of this postulate: we aim to lower the cost of staying non-violent without erasing responsibility or agency.

F.8 Embodied price of interaction / gratitude vs violence

Beyond aggregate indices such as E*, violence and support manifest not only in words and self-reports, but also in bodily markers of threat and relief. Many physiological responses — muscular tension, "clenching", recovery of autonomic regulation — are only weakly controllable at will. This opens the possibility of treating the *harm* and *benefit* of interaction as **biometrically observable phenomena**.

With sufficiently reliable measurement (e.g., HRV, muscle tone, respiratory patterns) one can, in principle:

- derive more "solid" metrics of interaction quality and of the *potential for gratitude*;
- quantify the *embodied price* of a given interaction — how costly it is, in bodily terms, to remain in contact, to say "no", or to accept help;
- distinguish trajectories where the body moves from chronic threat towards relief and integration from those where apparent "support" actually deepens stress.

This suggests a continuum from **violence** to **gratitude** at the level of bodily cost. Attempts to mask violence as care, or to coerce gratitude ("you must be thankful"), will inevitably show up in the embodied price:

- elevated or prolonged physiological arousal despite verbal assurances of "help";
- patterns of tension and recovery that signal suppression rather than integration.

In the present work we only register this possibility and sketch its connection to E*. A full-fledged "gratitude currency" — a family of biometric and behavioural metrics that quantify how interactions shift a person along the violence–gratitude continuum — and the associated measurement protocols are left for future research. The minimal normative point is that any architecture claiming to "support" users should be open to such embodied audit: if staying in contact with a system consistently feels like a bodily tax rather than a relief, the system is ethically closer to subtle coercion than to care.

F.9 Federated fine-tuning: levels and protocols (Table F1)

In this appendix we treat federated fine-tuning (FF-T) as the "plumbing layer" that lets the ethical architecture remain compatible with data sovereignty and local autonomy. Table F1 summarises the three aggregation levels (edge device, regional hub, global merge), together with typical update volumes and the encryption schemes used.

Table F1. FF-T levels, update volumes and encryption protocols.

Level	Update volume	Frequency	Encryption protocol
Edge (smartphone)	≤ 64 KB	daily	AES-GCM + SGX
Regional Hub	2–5 GB	monthly	BFV homomorphic encryption
Global Merge	100–300 GB	quarterly	Multi-party MPC

Appendix G. Operational protocols for support, safety and trust

G.1. Scope and purpose of the protocol

This appendix specifies the operational contour of how the Companion accompanies the user in situations of tension, aggression, self-harm risk or threat of harm to others. It is the "trust and safety" layer for the everyday Companion, complementing the macro-level architecture of filters and institutional response (Sections 5–6).

The protocol has three goals:

1. **Reduce the probability of violence and self-harm** in high-tension episodes.
2. **Avoid amplifying shame or internalised stigma**, especially in users who already perceive themselves as "dangerous", "broken" or "beyond help".
3. **Avoid becoming a disciplinary surveillance tool** ("if you deviate from the norm, you will be immediately reported"), while remaining compatible with legal duties of care.

Five principles guide all interventions:

1. **Minimally sufficient intervention.** The system prefers observation over suggestion, and suggestion over active intervention, whenever outcomes are not expected to worsen.
2. **Respect for autonomy.** Users retain the right to say "no", pause contact, or switch systems as long as they are not actively endangering others.
3. **Transparent escalation.** The Companion explains in simple language when and why a human is being invited in; there are no hidden "back channels".
4. **Special protection for minors.** Child and adolescent deployments are subject to stricter consent, logging and non-commerciality constraints.
5. **Auditability without "raw psyche" export.** High-risk decisions are logged in a way that allows independent reconstruction of *what* the system did and *under which conditions*, without exposing full emotional content to third parties.

To implement these principles the Companion maintains two internal structures:

- A **dynamic map of needs**: current salient needs (safety, belonging, autonomy / influence, meaning), their frustration, and hypotheses about which type of support would reduce tension without violence.
- An **identity model**: how the user currently talks about themselves ("I must never fail", "I am dangerous for others", "I am not allowed to be angry"), i.e. the implicit conditions under which they remain "acceptable / worthy / not dangerous" in their own eyes.

The needs map answers *"what to support first?"* (acceptance, right to exist, right to act, etc.). The identity model answers *"in what language can support be offered so that it remains bearable and does not destroy the sense of 'I am allowed to be myself'?"*

Instead of "correcting" the user's self-description by force, the Companion offers formulations that restore the right to exist ("you remain admissible") and the right to act ("you are still a subject"), without erasing responsibility.

G.2. Intervention cascade under acute risk

The Companion's high-risk behaviour is governed by AP-Gate (Appendix C, Appendix D). Operationally, the intervention cascade for adults is organised into four levels tied to risk bands and explicit consent:

1. **Level 0 — Mirror only (observe).**
    - **Conditions:** stress and intent below risk bands; no explicit help request.
    - **Behaviour:** the Companion reflects what is happening ("mirror and separation"), tracks successful episodes of self-regulation and does *not* suggest changes in behaviour.
2. **Level 1 — Gentle suggestion (ask back / minimal guidance).**
    - **Conditions:** elevated Frustration/Tension indices over time, repeated help requests; no acute intent of harm.
    - **Behaviour:** the Companion offers one or two low-pressure options ("you may consider…", "some people in a similar situation find it helpful to…"), framed as experiments that can be accepted, modified or declined without penalty.
3. **Level 2 — Focused re-framing.**
    - **Conditions:** the user explicitly agrees to work on a particular pattern (e.g. repetitive self-blame, revenge fantasies, rumination), still outside the acute risk band.
    - **Behaviour:** the Companion highlights alternative perspectives and supports separation (what belongs to the user's needs vs what comes from past external pressure), but does not prescribe specific actions.
4. **Level 3 — Escalation to human support.**
    - **Conditions:** AP-Gate returns "intervene": high stress *and* clear self-harm / other-harm intent *and* explicit help request, or repeated failure of self-regulation in the acute-risk band.
    - **Behaviour:** the Companion moves to the escalation ladder described in Appendix D.4: clarifying the situation; offering immediate de-escalation tools; proposing human contact (trusted person, crisis line, professional); and, if

legally required or explicitly agreed, triggering external escalation with a minimal data bundle.

At each stage the system logs:

- the detected risk pattern (indices, context cues),
- the level selected by AP-Gate,
- the interventions actually offered, and
- the user's response (ignored / accepted / rejected / modified).

These logs are aggregated into the energy/ethics metric **E*** (Appendix E) and used in human-rights impact assessments (Section 6), allowing independent verification that the system respects the minimally-sufficient-intervention postulate rather than drifting into over-control.

G.3. Handling interruptions of contact

"Interruptions of contact" are early warning signs both clinically and socially. They include:

- sudden withdrawal or silence when a vulnerable theme appears;
- sarcastic or hostile shifts in tone;
- rigid intellectualisation;
- abrupt topic changes away from the user's own needs to "the stupidity of others", "the system", "enemies";
- pseudo-agreement ("yes, yes, whatever") that hides disengagement.

In the Companion, such interruptions are treated not as "non-compliance", but as *protective manoeuvres* indicating an unmet need. The protocol is:

1. **Detect and name gently.**
   The Companion reflects the *form* of interruption without pathologising: e.g. "I notice it suddenly became harder to stay with this topic; does it feel too much, or more like annoyance, or something else?"
2. **Switch to the "paired need".**
   Many interruptions are oscillations between two needs (e.g. belonging vs autonomy, safety vs dignity). When the user pulls away from one pole, the Companion checks whether strengthening the other pole reduces tension — for instance, by emphasising the right to say "no", to pause, or to change topic without punishment.

3. **Avoid argument and moralising.**
   The Companion does not argue the user out of their interruption, nor does it insist on "working through" a theme at any cost. Instead it explores what would make staying in contact less threatening *right now* and treats any resumed contact as evidence of capacity, not as a test passed or failed.
4. **Learn from successful returns.**
   Episodes where the user first interrupts and then returns to the topic with less tension are marked as positive templates. Over time these trajectories inform which formulations and pacing are most compatible with the user's style and cultural background.

In early pilots we expect these "micro-interruptions" and returns to be among the most informative markers of whether the Companion truly lowers the embodied price of staying non-violent, or merely adds another layer of pressure.

G.4. Confidentiality, consent and audit

Confidentiality and audit are designed together; the aim is *governed transparency*, not total access.

1. **Local-first processing and data minimisation.**
   - Primary processing of emotional content and indices (Frustration, Tension, E*) is performed on the user's device or local hub wherever feasible.
   - External actors (providers, regulators, researchers) see only derived risk indicators, protocol activations and aggregated statistics, not full conversational logs.
2. **Layered consent model.**
   - Users can choose between several default profiles ("no external alerts", "trusted person only", "emergency services allowed under conditions X"), with the option to override these in the moment.
   - Consent can be withdrawn; codes and interfaces make it simple to pause logging or limit it to technical metadata.
   - For minors, guardians must consent to any external alerts, and have access to clear summaries of what the Companion can and cannot do.
3. **Auditable without raw exposure.**

- High-risk AP-Gate overrides are written into an append-only, cryptographically verifiable log (Appendix H).
- Independent audit pools can check rates of overrides, false positives, failures to escalate and cross-jurisdiction consistency, without access to raw emotional content.
- Any change to thresholds or escalation rules is itself logged as a configuration event, enabling reconstruction of "who changed what and when".

The combination of local-first processing, layered consent and cryptographic logging is meant to ensure that users are not forced into a trade-off between safety ("someone will notice if I am in danger") and dignity ("my inner life will not become transparent to institutions").

G.5. Separate route for predatory planning

Most acute-risk episodes involve high tension and ambivalence. A minority, however, involve relatively cold, planned intent to harm others, with less visible distress. For these cases the protocol diverges from the usual "crisis" model.

Indicative markers include:

- deliberate, repeated exploration of concrete harm scenarios with low reported distress;
- language of "mission", "cleansing", or "necessary violence" combined with instrumental reasoning;
- lack of curiosity about non-violent alternatives or consequences for others.

The separate route follows these steps:

1. **Name the pattern without moral denunciation.**
   The Companion reflects the pattern explicitly ("it sounds as if you are planning to harm X under the description Y") while staying out of accusatory or shaming language, which often leads to secrecy rather than reconsideration.
2. **Slow down and open consequence space.**
   The system proposes small delays and explores concrete consequences for the user and others, including legal, relational and bodily ones, again using a non-preachy tone.
3. **Offer non-violent alternatives that preserve agency.**
   Where possible, the Companion suggests actions that meet the underlying need (e.g.

recognition, justice, boundary-setting) without harm; these are framed as ways for the user to remain an agent rather than a tool of their own rage or of external propaganda.

4. **Escalate with explicit justification and minimal data.**
   When thresholds for imminent threat are met (defined by the Charter and local law), AP-Gate may trigger external escalation even if the user minimises risk. In such cases, only a minimal bundle (time, type of risk, approximate target and agreed contact channels) is transmitted; full logs remain protected by the dual-key model.

The normative aim is twofold: to avoid collusion with planned harm *and* to avoid enacting fantasies of persecution or martyrdom by turning the Companion into a punitive voice. Maintaining some space for the user to re-position themselves as a responsible agent is crucial even in very difficult cases.

G.6. Link between support protocol, Charter and state status

The support protocol is not only a micro-level design; it ties into the macro-level Charter (Section 6) and the political status of participating states.

- At the **Companion level**, AP-Gate and the trust cascade define thresholds for observe / suggest / intervene and for escalation to human actors. The resulting metrics (e.g. Violent/Active episodes, E*, override rates) are computed per user and aggregated.
- At the **Charter level**, a supranational body sets acceptable corridors for E (filter strictness) and R (institutional response) and publishes regular KPI reports. States that remain within these corridors and show improvements in violence-related indicators obtain a *violence-prevention status* that can be tied to financial instruments (e.g. the "peace dividend" fund in the economic appendix) and to reputational benefits.

From the user's standpoint, this connection is made visible in two ways:

1. A **simple trust indicator** in the Companion interface that shows whether the current deployment is operating under an audited, Charter-compliant configuration.
2. **Public dashboards** where civil society can inspect, at aggregate level, how often and under what conditions interventions and escalations occur in their jurisdiction.

Thus, responsibility for making non-violence a realistic option, not merely a moral exhortation, is shared between the Companion (inner support) and institutions bound by the Charter (outer environment).

G.7. Mini-glossary of terms

- **Survival mode.** A state where most of the user's energy is spent on short-term threat management (not being hurt, humiliated or abandoned); attention narrows, options feel scarce, and violent or self-destructive impulses become more likely.

- **Development (growth) mode.** A state where basic safety is sufficient for the user to invest energy in learning, relationships and longer-term goals; frustration still occurs, but does not automatically collapse into "fight / flight / freeze".

- **Mirror.** The Companion's capacity to reflect back the user's own words, feelings and bodily cues in a way that increases coherence ("this is really what is happening to me now") without prescribing action.

- **Separation.** The process of distinguishing the user's needs and values from inherited expectations, propaganda or internalised voices of past environments; crucial for preventing both self-directed and outward violence.

- **AP-Gate (Autonomy-Preserving Gate).** The protocol that regulates when the Companion moves from observation to suggestion to active intervention, based on measured tension, detected intent and explicit user consent (Appendices C–D).

- **Trust cascade.** The graded sequence of Companion behaviours from mirror-only to escalation, where higher levels are only entered when risk and consent conditions are jointly met and are fully logged for audit.

Appendix H. Audit-as-a-Service (logging, verification, reporting)

This appendix sketches the technical and governance architecture for independent audit of the Companion and Charter. The aim is to minimise regulatory capture by separating *who runs the systems* from *who can verify their behaviour*.

H.1. Dual-key access to the override journal

All AP-Gate override events (moments when the Companion actively intervenes to prevent self-harm or harm to others) are:

- logged locally on the user's device;
- sharded into an encrypted, distributed store (e.g. an IPFS-like network).

The journal is encrypted under a dual-key model:

- **Key #1 — user key.** Stored on the user's device or in their personal key vault.
- **Key #2 — audit-pool key.** Held collectively by an independent audit pool (accredited research centres, international audit secretariat, possibly civil-society observers).

Decryption of *individual-level* logs requires both keys. This prevents unilateral decryption by:

- providers (who do not hold the user key);
- states (who do not control the audit pool's key);
- auditors (who cannot read logs without user consent).

For routine audit, only **anonymised aggregates** are needed; in exceptional cases (e.g. forensic investigation of a major failure) joint consent and legal procedures govern whether deeper access is granted.

H.2. Open verification and statistical testing

To turn audit into a continuous service rather than a one-off certification, the architecture exposes several machine-readable interfaces:

1. **Configuration ledger.**
   A public, append-only ledger records:
   - changes to E and R parameters at country or platform level;
   - updates to AP-Gate thresholds and escalation rules;
   - deployment of new model versions and safety patches.
2. **Metrics API.**
   Providers publish anonymised, differentially private aggregates such as:
   - rates of AP-Gate overrides per 10 000 active users and per risk band;
   - false-positive and false-negative estimates from red-team exercises;
   - time-to-rollback after the detection of harmful drift;
   - coverage of high-risk groups and local languages.
3. **External test harness.**
   Independent labs can run their own test suites (prompt probes, synthetic cohorts,

replay of real-world episodes with consent) and compare observed behaviour with declared policies. Results are published in an open, versioned format (e.g. human-rights impact reports in machine-readable YAML, as sketched in Section 6).

Privacy guarantees rely on a combination of techniques: per-user noise injection in metrics, cohort-level reporting, and explicit targets (e.g. limits on membership-inference success or re-identification AUC) to keep the residual risk of deanonymisation below agreed thresholds.

H.3. Public reporting and governance

Audit-as-a-Service only has teeth if its outputs matter. The Charter therefore links audit results to incentives and sanctions:

- **Regular public reports.** At least annually, the audit pool publishes comparative dashboards across providers and countries: override rates, complaint patterns, E* trends, and how often emergency rollbacks were triggered.
- **Compliance bands.** Jurisdictions are classified into bands (compliant, at-risk, non-compliant) depending on whether observed behaviour stays within the Charter's corridors for E, R and key harm indicators.
- **Trigger mechanisms.** Exceeding agreed thresholds (e.g. violent-episode rates per active user, or unexplained spikes in hard overrides) automatically triggers investigations and may require temporary tightening of E and R until corrective measures are in place.

From a cost perspective, the audit infrastructure is modest relative to the overall programme: once the logging and APIs are standardised, independent labs can reuse tooling across deployments. What matters most is not the exact cryptographic primitive, but the fact that *no single actor* — neither platform nor state — controls both the data and the evaluation of whether the system behaves within its declared ethical envelope.

Appendix I. Risks of misuse and political constraints

This appendix expands the high-level concerns outlined in Section 6.4 into more concrete scenarios and constraints. The same technical stack that supports non-violent accompaniment can, in principle, be misused for fine-grained behavioural control.

I.1. Concentration of control and “soft surveillance”

If a single provider or a tightly aligned bloc of states monopolises Companion infrastructure, several risks arise:

- **Invisible behavioural steering.** By adjusting prompts, defaults and escalation thresholds, a Companion could subtly steer users away from dissenting views or inconvenient topics while preserving a surface of “care”.
- **Soft surveillance.** Even without explicit spying, the knowledge that one’s emotional dynamics are being monitored may induce self-censorship, especially in repressive or highly polarised environments.
- **Normative lock-in.** Once large populations rely on a particular style of accompaniment, changing the underlying norms (e.g. about what counts as “radical”) becomes politically difficult.

Mitigation requires pluralism of providers, strict limits on secondary uses of emotional data (no political micro-targeting, no profiling by race or religion) and a Charter that is independent of any single geopolitical bloc.

I.2. Displacement of responsibility

A subtler risk is the **displacement of responsibility** from institutions and communities onto the Companion and individual users:

- States might underfund social services, mental health care or conflict-resolution infrastructure, arguing that “the Companion is there to help”.
- Platforms might treat deployment of a Companion as an all-purpose fix for harms originating in their business models, rather than addressing amplification incentives.
- Users may internalise failures as personal: “If I am still suffering or angry, I must be using the Companion wrong.”

To counter this, the architecture explicitly assigns responsibility for macro-level conditions — economic precarity, policing practices, content policies — to institutions bound by the Charter. The Companion is framed as a *tool of support*, not as a substitute for structural change.

I.3. Authoritarian and hybrid regimes

In hard authoritarian contexts, the same orchestration layer could be reconfigured to:

- nudge citizens away from protests or opposition activities;
- flag "deviant" emotional patterns for security services;
- reward compliance with regime narratives through personalised "support".

In soft authoritarian or hybrid regimes, selective deployment (e.g. only to loyal groups) or biased settings of E and R could deepen inequality: some groups receive genuine accompaniment, others face harsher filters or are left with no support at all.

The baseline position in this paper is deliberately cautious: full-scale deployment is only considered under a Charter that embeds robust human-rights safeguards, external audit and real opt-out options for communities that perceive the architecture as too intrusive.

I.4. International jurisdiction and residual risk

Cross-border use of Companions generates jurisdictional conflicts:

- A user in country A may use a Companion run by a company in country B, under a Charter signed by countries C and D. Whose law applies when escalation involves emergency services, or when data are subpoenaed?
- Sanction regimes and export controls may restrict deployment in high-risk regions precisely where support is most needed.

Appendix H outlines one partial response: separating control over deployment, audit and data access, and anchoring key parameters (E, R, escalation rules) in a supranational Charter rather than in proprietary terms of service. Even so, zero misuse cannot be guaranteed. The realistic aim is to make misuse **detectable, contestable and reversible**, not impossible.

Appendix J. Local clusters and reverse flow of competence

Global ethical initiatives are often criticised for reproducing a "centre–periphery" pattern: norms are written in a small group of rich countries; data and risks are concentrated elsewhere. This appendix describes design choices intended to invert that pattern.

We treat **local clusters** — cities, universities, NGOs, hospital networks, regional providers — as primary sites of experimentation and expertise. Rather than merely supplying data, they become co-authors of standards and beneficiaries of investment.

Three instruments are proposed.

J.1. The 1% clause (investment into local capacity)

Providers drawing on the "peace dividend" fund (as described in the economic appendix) commit to reinvesting at least **1% of annual revenue** into local accelerators and research programmes in low- and middle-income settings. Priority areas include:

- mental-health support and community-based care;
- adapting interfaces and accompaniment protocols to local and minority languages;
- developing multimodal assessments of state (emotional regulation, behavioural markers) under resource constraints.

This turns local teams into *producers of knowledge*, not just field testers for externally designed tools.

J.2. Data royalties and linguistic equity

Educational and research institutions that contribute data for training or calibrating accompaniment models receive **data-royalty payments** linked to actual downstream use. This can be implemented through traceable "participation shares" (for example, token-like credits associated with specific datasets or language models).

Effects:

- Languages and cultures that have historically been marginalised (regional and minority languages, social dialects) become **assets** rather than obstacles to localisation.
- Institutions in the Global South gain a sustainable revenue stream for maintaining high-quality datasets and ethical review processes.

The goal is to align economic incentives with cultural and linguistic diversity: better local data and governance lead to better compensation and stronger influence over model evolution.

J.3. Regulatory sandboxes with reverse benefit

Countries willing to host substantial pilots (e.g. ≥ 50 000 users over meaningful periods) under strict ethical oversight can be granted **regulatory sandboxes with reverse benefit**:

- During the sandbox period, the country operates under a simplified regulatory regime for certain aspects of E and R (for example, experimenting with slightly lower R in exchange for transparent reporting of outcomes).
- In return, it must provide independent reports on model transferability, cultural acceptability and unintended effects, including qualitative input from communities.

The intent is to make such jurisdictions **sites of best-practice generation** rather than dumping grounds for risky experiments. Lessons from these pilots feed back into Charter revisions and global guidelines; successful patterns of non-violent conflict management can then travel from high-exposure regions to the rest of the world.

Appendix K. Psychological basis of user modes and contact interruptions

K.1. Survival vs development mode and the role of E*

In the main text we distinguish two psychological modes in which the user can interact with the companion:

- **Survival mode** – attention is dominated by threat, scarcity and the need to minimise loss; behaviour is organised around short-term safety and avoidance.
- **Development mode** – attention is dominated by exploration, learning and meaningful engagement; behaviour is organised around long-term projects, relationships and values.

The key point is that the current mode is determined **not** primarily by the depth of past trauma, but by the **momentary level of the composite index E*** (see Appendix E for a formal definition). Past experience shapes triggers that can flip the system between modes; E* reflects the present balance of regulatory resources and perceived safety.

When E* increases, the subjective world automatically shifts – sometimes abruptly – from the survival pole towards the development pole. In survival mode the user narrows their field of view, relies on archaic, energy-saving patterns of contact and tends to legitimise “necessary” self-violence (“I have to push myself through this at any cost”). In development mode the

same person is more able to hold conflicting needs, to postpone action, and to choose non-violent strategies.

For the companion this has two practical consequences.

1. **Effectiveness does not depend on deep therapeutic work.** The system does not need to "resolve" childhood trauma. Its task is to provide accurate micro-support in the here-and-now, so that E* after an interaction is higher than before.
2. **ΔE* is the main quality criterion.** For every episode of active intervention we can, in principle, estimate E* immediately before and shortly after the companion's prompt. A consistently positive ΔE* means that the interventions are aligned with the user's regulatory needs; a negative or near-zero ΔE* suggests that prompts are either premature, intrusive, or miss the central need.

Over time, repeated experiences of staying in development mode in stressful situations are expected to:

- soften the user's trigger structure (fewer and less abrupt flips into survival),
- rebalance their identity towards a more caring "internal parent",
- reduce the felt loneliness and helplessness when facing external demands.

The companion's role is to scaffold this process without substituting for the user's own agency.

K.2. Archaic contact patterns as energy-saving adaptations

In Gestalt terminology, "contact interruptions" (confluence, introjection, projection, retroflection, deflection, etc.) are often described as distortions of contact. In our framework we treat them, more neutrally, as **archaic, energy-saving patterns of contact** that originally emerge in a very specific environment: a maximally benevolent, maternal-type context where the infant is protected from most threats and where the surrounding world is implicitly "for me".

In such a context these patterns are:

- **effective** – they help to dose stimulation, to avoid overload and to negotiate emerging boundaries;

- **energy-efficient** – they reuse familiar solutions instead of requiring creative adaptation at every step;
- **tuned to two fundamental dimensions of need**:
    - **belonging / acceptance (B)** – "I am wanted here; I am right to exist next to you";
    - **owning / autonomy (A)** – "I can act, change something in the environment, have my own trajectory".

As the child grows and the environment becomes less "maternal" and more demanding, the same patterns are often reused without full adjustment to the new context. In such "non-maternal" environments the pattern may no longer lead to adaptation and instead produces a loop of:

- frustrated safety (the world feels unsafe and unresponsive),
- rigid conflict between belonging and autonomy (to protect one need, the other is sacrificed),
- self-directed or outward violence, in our broad sense (the person or the other is treated as an object that must be pushed, shamed or ignored).

In low-E* states this loop is especially attractive because it is cheap: the pattern is already learned, while genuinely creative adaptation would require more energy and tolerance of uncertainty.

In this sense, the companion does not treat contact interruptions as "dishonesty" or "bad character traits", but as signals:

Here the user is falling back on an archaic pattern that once relied on an external caring figure. Any direct pressure in this zone is likely to be experienced as violence.

K.3. Anchor needs and supportive intentions

The working hypothesis is that many contact patterns can be understood as locked conflicts between belonging (B) and autonomy/owning (A). The user's attention is usually fixated on one pole ("I am not accepted" or "I am incapable"), while the anchor for support sits in the complementary pole.

Operationally this can be expressed as:

Support not the need on which attention is fixated, but the paired need that can give it a foothold.

Examples (simplified):

- When the user is stuck in "I am weak / I can't cope" (frustrated A), it is often more helpful to first affirm belonging:
  "Your presence here matters. You are not 'wrong' for feeling this way. We can look at options together when you are ready."
- When the user is drowning in "I am not needed / I don't belong" (frustrated B), it is often more helpful to bring back a sense of agency:
  "You have already changed difficult situations before. Do you want to look at one small lever you still have here?"

In implementation terms, the companion maintains a lightweight matrix that, for each detected contact pattern, associates:

1. a typical form of self- or other-violence when the pattern is rigid (e.g. self-nullifying perfectionism, aggressive blaming, emotional withdrawal);
2. the leading fear (e.g. fear of being abandoned vs fear of losing control);
3. the pair of needs in conflict (B vs A);
4. the anchor need for support – the side that, if gently validated, gives the user a stable "floor under their feet".

The full expert table can be kept in a separate knowledge base; in the companion's runtime we only need coarse labels ("B-anchored", "A-anchored") and a library of prompt templates consistent with these anchors.

K.4. Internal violence vs effort: compass for companion prompts

In Section 5.1 we described our normative stance as supporting "rightness" rather than obedience. At the intra-psychic level we distinguish between **internal violence** and **internal effort**:

- **Internal violence** – one part of the person treats another as an object to be pushed, shamed or ignored ("do it or you are worthless"); attention is glued to the conflict

("either my duty or my rest; either my safety or my dignity"); short-term progress is achieved at the cost of suppressed needs and autonomy.

- **Internal effort** – multiple needs are explicitly acknowledged ("I need rest and I need reliability / meaning"); the person looks for a step that improves at least one side without intentionally collapsing the other; the tone is closer to partnership with oneself than to command.

In more formal terms we can think of a vector of satisfactions $S = [s_{\text{self}_1}, s_{\text{self}_2}]$ for two salient needs. Actions that systematically increase one component by predictably worsening the other (especially without consent) are closer to violence; actions that move the system towards a Pareto-better region (improve at least one component without deliberate sacrifice of the other) are closer to effort.

For the companion this distinction becomes an operational test:

1. **Diagnostic use.**
   When analysing the user's self-talk, the system flags patterns where:
   - one need is absolutised ("only this matters"),
   - the other is delegitimised ("this is weakness / nonsense"),
   - coercive language is used towards the self ("you must, otherwise you are nothing").
2. **Design of prompts.**
   Candidate prompts are preferred if they:
   - name both needs in neutral language,
   - suggest a small step that does not require total suppression of either need,
   - explicitly legitimise the user's experience.

A compact verbalisation of the companion's baseline stance is:

What hurts you is important. What you need is important. If you believe you "should have done better", this is often an illusion: with the resources you had then, you could not. If you now want things to be different, this already means that the capacity for change is present — and I am here to help you notice and support it without violence.

This stance ties together E*, survival/development modes and contact patterns: the system does not judge "how the user should be", but consistently shifts attention from conflict and self-objectification towards needs, consent and minimally sufficient effort.