



Self++: Co-determined agency for human–AI symbiosis in extended reality

Thammathip Piumsomboon*

School of Product Design, University of Canterbury, Christchurch 8041, Canterbury, New Zealand.

***Correspondence to:** Thammathip Piumsomboon, School of Product Design, University of Canterbury, Christchurch 8041, Canterbury, New Zealand. E-mail: tham.piumsomboon@canterbury.ac.nz

Received: February 17, 2026 **Accepted:** April 24, 2026 **Published:** April 29, 2026

Cite this article: Piumsomboon T. Self++: Co-determined agency for human–AI symbiosis in extended reality. *Empath Comput.* 2026;2:202604. <https://doi.org/10.70401/ec.2026.0021>

Abstract

Self++ is a conceptual design framework for human–Artificial Intelligence (AI) symbiosis in extended reality (XR) that preserves human authorship while still benefiting from increasingly capable AI agents. Because XR can shape both perceptual evidence and action, apparently ‘helpful’ assistance can drift into over-reliance, covert persuasion, and blurred responsibility. Self++ grounds interaction in two complementary theories: Self-Determination Theory (autonomy, competence, relatedness) and the Free Energy Principle (predictive stability under uncertainty). It operationalises these foundations through co-determination, treating the human and the AI as a coupled system that must keep intent and limits legible, tune support over time, and preserve the user’s right to endorse, contest, and override. These requirements are summarised as the co-determination principles (T.A.N.): Transparency, Adaptivity, and Negotiability. Self++ organises augmentation into three concurrently activatable overlays spanning sensorimotor competence support (Self: competence overlay), deliberative autonomy support (Self+: autonomy overlay), and social and long-horizon relatedness and purpose support (Self++: relatedness and purpose overlay). Across the overlays, it specifies nine role patterns (Tutor, Skill Builder, Coach; Choice Architect, Advisor, Agentic Worker; Contextual Interpreter, Social Facilitator, Purpose Amplifier) that can be implemented as interaction patterns, not personas. The contribution is a role-based conceptual map that generates testable design propositions for XR-AI systems that grow capability without replacing judgement, enabling symbiotic agency in work, learning, and social life and resilient human development.

Keywords: Extended reality, human-AI symbiosis, human-AI teaming, self-determination theory, free energy principle, co-determination, augmented agency, interaction design, trustworthy AI

1. Introduction

Seen in a longer arc, the present moment is a new chapter in human cognitive extension. Early visions of human–computer integration already anticipated a deep mind–machine symbiosis. In 1960, Licklider described “man–computer symbiosis” as an interactive partnership where humans and computers complement each other’s strengths^[1]. Shortly after (1962), Engelbart proposed augmenting human intellect to boost problem-solving through technology^[2]. Later theories formalised this relationship: the extended mind hypothesis argued that artifacts can become literal parts of cognition^[3], while distributed cognition emphasised that thinking often spans people, tools, and environments^[4]. Empirical work supports parts of longstanding worries about cognitive erosion, including the “Google effect”^[5], inflated perceived knowledge from internet access^[6], and weakened spatial memory from heavy global positioning system (GPS) reliance^[7]. At the same time, cognitive science and Human–Computer Interaction (HCI) stress that “thinking with things” can improve reasoning^[8], and that distributing cognitive load can enable more complex problem-solving^[9].

We are accelerating toward more autonomous productivity, driven by virtual agents and embodied robots. Toolchains, platforms, and organisational agent stacks are being optimised for throughput, reliability, and delegation, first under human instruction and, plausibly, under higher-level supervisory agents. This trajectory sharpens an old question: how do we gain the benefits of delegation without losing authorship? Much current work focuses on catastrophic risks, from extreme harm like misalignment to subtler manipulation and societal subordination^[10,11]. Alongside these frontier concerns is a nearer failure mode: artificial intelligence (AI)-mediated convenience can erode careful reasoning, making people less critical and more easily persuaded^[12]. This predates modern generative AI, but today’s systems are more fluent, more personalised, and more persuasive at scale.

This acceleration represents a major expansion of the human cognitive niche. Evolutionary accounts suggest humans gained advantage through causal reasoning, tool-making, and cooperative action rather than biological arms races^[13]. Humans externalised



thinking into artifacts and social systems, expanding intelligence through culture. This tendency is captured by the notion that we are “natural-born cyborgs”^[14] living in co-evolution with our technologies^[15]. Today, the niche expands again through extended reality (XR) and AI. XR extends perception and presence in mixed and virtual environments, while AI externalises reasoning by perceiving and acting alongside us. The result is a tightly interwoven human-machine ecology in which cognitive strategies are co-generated.

This expansion also carries costs. The complexity and volatility of an XR-AI ecosystem can create an “entropy challenge”: unpredictable stimuli and shifting agent behaviours that outstrip our capacity to maintain coherent models^[15,16]. The mismatch can manifest as information overload, attentional fragmentation, and blurred agency, where users may be unsure where their intentions end and the system’s suggestions begin. These symptoms point to design failures in how autonomy, competence, and social connection are protected under delegation. Emerging phenomena make this gap visible, including split-attention demands in XR workflows^[17], identity confusion as agents become more human-like^[18,19], and misaligned persuasion where simulation success does not translate into real behaviour change^[20]. An “AI loneliness trap” may also emerge, where convenient synthetic companionship gradually displaces human relationships for some users^[21].

If AI is the latest extension of cognition, synergy is not automatic. Automation research has long catalogued challenges such as common ground, trust calibration, and complacency^[22]. A large meta-analysis suggests that human-AI teams often fail to outperform the better of the human or AI alone^[23,24], implying that effective teaming must be deliberately designed with interaction mechanisms that respect human cognition and shared agency. Recent work reinforces that the most accurate AI is not always the best teammate^[25]. Human-computer integration research argues for a deeper intertwining where systems adapt to the user’s state in real time and users fluidly delegate and reclaim control^[26]. XR makes this tangible by placing intelligent support in the user’s perceptual and social space. Augmented reality (AR) enables in-situ augmentation^[27,28], including social interventions like calming virtual companions^[29], while virtual reality (VR) provides rich environments for studying collaboration dynamics^[30,31]. Together, these platforms make the hybrid niche observable and designable, strengthening the case for co-determined agency as a design primitive.

Self++ responds to this agency problem by grounding design in two complementary foundations: Self-Determination Theory (SDT; autonomy, competence, relatedness) and the Free Energy Principle (FEP; predictive stability under uncertainty). In simple terms, SDT specifies what must be preserved for flourishing, while FEP explains why the pressure intensifies as environments become more volatile and mediated. We operationalise these foundations through co-determination, allowing users the right to endorse, contest, and override assistance. We summarise these requirements as three co-determination principles (T.A.N.): Transparency, Adaptivity, and Negotiability.

These pressures matter deeply to me as an educator and parent. Self++ is not meant to diminish AI’s utility or reject autonomy as a design direction. Instead, it offers a framework for human-AI systems that preserve the benefits of intelligent support while strengthening user agency, competence, and social integration. Self++ organises augmented agency into three concurrently activatable overlays and nine role patterns spanning sensorimotor support, deliberative decision support, and longer-horizon social and purpose alignment (Figure 1). We present Self++ as a theoretically grounded perspective rather than a validated specification. The framework is intended to organise an emerging design space, generate testable propositions, and provide conceptual tools for researchers and practitioners navigating the complex territory of human-AI collaboration in XR. Its value lies in making this territory more tractable for systematic investigation, not in prescribing final solutions. Across all role patterns, T.A.N. functions as a practical safeguard, so uncertainty regulation supports human development rather than bypassing it.

2. Background

2.1 Theoretical foundations: Self-determination and free energy

As noted in the Introduction, the emerging XR-AI ecology raises the stakes for supportive AI design. Self++ is grounded in two complementary pillars: SDT^[32] and the FEP^[16]. Together, they explain why autonomy, competence, and relatedness matter under intelligent mediation, and how systems might support them without eroding agency. SDT identifies three basic psychological needs (autonomy, competence, and relatedness) as essential for motivation and well-being^[32]. Autonomy is volitional, self-endorsed action; competence is efficacy and skill growth; relatedness is connection and belonging. When these needs are supported, people show stronger intrinsic motivation, learning, and well-being; when they are chronically frustrated, disengagement and poorer performance follow.

The Free Energy Principle complements SDT by explaining why these needs become harder to protect as environments grow more complex. Predictive processing accounts formalise perception and action as approximate Bayesian inference aimed at minimising surprise^[33]. Friston’s FEP generalises this: biological systems maintain integrity by minimising free energy, closely related to long-run prediction error^[16]. The brain updates internal models to keep sensory input within expected bounds; as environments become more volatile, the computational and attentional cost of this stabilisation rises. Because human predictive models evolved for relatively stable ecologies, they can be mismatched to fast-changing, algorithmically mediated XR, producing sustained prediction errors experienced as stress, confusion, or overload^[15].

Self++ bridges SDT and FEP by treating autonomy, competence, and relatedness as stabilising conditions for predictive cognition: competence improves anticipatory control, autonomy protects against externally imposed goals that conflict with internal priors, and relatedness offloads uncertainty onto shared social models.

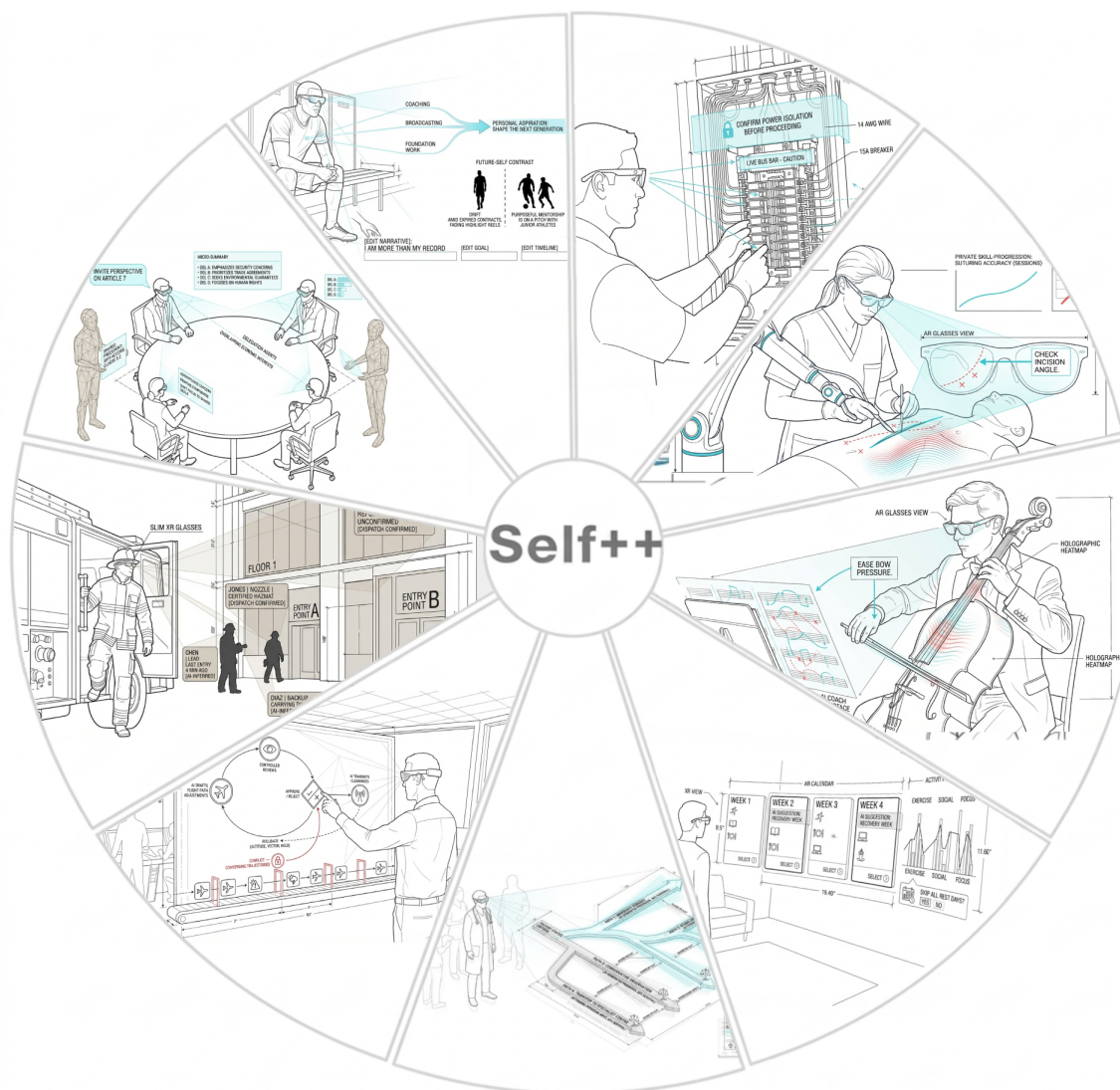


Figure 1. The nine Self++ role patterns organised across three concurrently activatable overlays, with co-determination principles (T.A.N.) scaling in strength with overlay scope and initiative. **Overlay 1 (Self): Competence support.** R1, *Tutor*: reduces novice uncertainty through a safe, learnable corridor (e.g., a trainee electrician receives anchored directional arrows, step gating, and ghosted hand exemplars through XR glasses while working on a residential electrical panel). R2, *Skill Builder*: calibrates and generalises skill through variability and augmented feedback (e.g., a training doctor receives real-time motion traces and a holographic accuracy heatmap overlaid onto a practice mannequin during a surgical procedure). R3, *Coach*: builds robustness under stress and supports self-correction (e.g., a cellist receives intonation feedback, fingerboard pressure heatmaps, and metacognitive prompts during a live performance, with social comparison replaced by private progression tracking). **Overlay 2 (Self+): Autonomy support.** R4, *Choice Architect*: shapes the decision context while preserving authorship (e.g., a person views a floating AR monthly calendar where recovery weeks are gently highlighted and a friction gate requests confirmation before overriding rest days). R5, *Advisor*: externalises deliberation by making counterfactuals and trade-offs inspectable (e.g., an ER doctor sees a branching holographic decision tree with uncertainty bands, survival-confidence estimates, and provenance badges distinguishing AI prognosis from attending physician input). R6, *Agentic Worker*: executes delegated tasks under a proposal-approval loop with rollback (e.g., an air traffic control shift manager oversees an AI-drafted routing queue where conflict items are flagged and rerouted back for manual handling, with any clearance reversible before transmission). **Overlay 3 (Self++): Relatedness and purpose support.** R7, *Contextual Interpreter*: makes identity, norms, and downstream impacts legible to reduce social surprise (e.g., a firefighter arriving at an incident sees AR-labelled crew roles, building entry points, and provenance badges distinguishing dispatch-confirmed from AI-inferred information). R8, *Social Facilitator*: improves human-to-human coordination and repair (e.g., diplomats at a round-table negotiation receive personalised, opt-in AR overlays including speaking-time balance, perspective-invitation prompts, and neutral micro-summaries of each delegation's position, while embodied virtual agents surface shared precedents as common ground). R9, *Purpose Amplifier*: supports long-horizon value coherence by making future trajectories legible and editable (e.g., a retiring athlete views a holographic value-map converging personal strengths toward an aspiration, with a future-self contrast between drift and purposeful mentorship and editable identity-narrative fields).

To operationalise these needs in interaction design, we propose *co-determination*. Rather than command-following tools or unilateral autonomy, codetermination treats human and AI as a coupled system that negotiates control to preserve psychological stability and minimise prediction error. We summarise this stance as three co-determination principles (T.A.N.):

- **Transparency:** From an FEP perspective, the system should minimise “hidden states.” If an agent’s reasoning or uncertainty is opaque, users cannot predict its behaviour, increasing anxiety and miscalibrated trust. Transparency supports accurate user mental models^[33].
- **Adaptivity:** From an SDT perspective, support should track the user’s changing *competence and relatedness conditions*. Static assistance becomes either intrusive (thwarting autonomy) or insufficient (thwarting competence) or socially mistuned as multi-party interaction evolves. Adaptivity provides scaffolding that intensifies or fades, or reconfigures to match learning, context, and the changing dynamics of social coordination^[34].
- **Negotiability:** Rooted in autonomy, users must be able to endorse, decline, or override system actions. Without negotiation or reversal, users lose authorship and volition^[32].

2.2 Philosophical and cultural influence

SDT and FEP explain how humans sustain motivation and stability under uncertainty, but they leave a prior normative question open: what kinds of selves are being shaped as perception, action, and social interaction become increasingly mediated by intelligent systems? This subsection adds a philosophical and cultural framing for why agency, interpretation, and social embeddedness must remain central and why T.A.N. becomes an ethical requirement.

Across many traditions, selfhood is relational and enacted through conditions rather than fixed or self-contained. In Buddhist philosophy, the self is an impermanent process (*anattā*) arising through interdependence^[35]. Māori ontology similarly foregrounds relational identity through *whakapapa* and frames well-being as sustained through balanced relationships among people, communities, and environment^[36]. Cross-cultural psychology likewise distinguishes relational and individual-centred models of self, showing that meaning, obligation, and autonomy are socially situated^[37]. These views do not reject autonomy; they recast it as accountable and context-sensitive.

Cognitive science offers a parallel account. Embodied and enactive approaches argue that sense-making emerges through ongoing organism–environment coupling, not detached internal reconstruction^[38–40]. The extended mind thesis similarly holds that cognition can span brain, body, artefacts, and social structures^[3]. If selves are enacted through tools and relationships, AI is not just an external utility; it helps shape the conditions through which identity and experience are repeatedly constructed. Self++, therefore, treats the social environment as a functional substrate for agency and understanding, not an optional dimension.

A further implication of this relational stance is that what counts as autonomy, competence, and relatedness, and how they are weighted, is not culturally uniform. Cross-cultural psychology has long shown that the boundaries of the autonomous self, the role of obligation in competent action, and the forms of belonging that sustain well-being differ substantially across individual-centred and relational self-construals^[37]. Indigenous frameworks such as Māori models of relational health foreground collective accountability and intergenerational connection as conditions for flourishing, not merely as a context for individual choice^[36]. Buddhist accounts treat selfhood as processual and interdependent, making the very notion of a bounded “autonomous agent” a convention rather than a ground truth^[35]. These differences are not peripheral to Self++; they are a primary reason why the framework adopts a procedural rather than substantive ethical stance. Self++ does not prescribe which values are correct or which model of selfhood is authoritative. Instead, it specifies interactional conditions (transparency, adaptivity, and negotiability) under which users and communities can recognise, reflect on, and act from their own endorsed commitments, whatever those commitments may be. T.A.N. is designed for moral and cultural plurality: transparency makes influence visible so it can be evaluated against local norms; adaptivity prevents the system from freezing around a single cultural default; and negotiability gives individuals and groups the power to contest, reconfigure, or refuse what the system surfaces and how. This procedural stance does not claim neutrality, because choosing what to make legible is itself a normative act, but it does ensure that such choices remain inspectable and revisable rather than hard-coded.

This relational stance is especially relevant for AI design when read alongside Dependent Origination (*pratītyasamutpāda*) and predictive processing. Dependent Origination holds that experience arises through interdependent causes and conditions. Predictive processing makes a structurally similar claim: perception is generated by hierarchical generative models that predict sensory input, so experience depends on the negotiation between signals and learned expectations^[16,33,41]. Prior work notes resonances between Buddhist accounts of interdependent experience and predictive approaches^[42]. Building on this, we map Dependent Origination to inferential perception: ignorance reflects model mismatch; formations shape priors; sense bases/contact sample evidence; and craving/clinging reflects the drive to resolve uncertainty, potentially hardening priors. The upshot is that perceptual qualities (e.g., a virtual object’s apparent “redness”), or qualia, are not intrinsic to stimuli, but emerge from relational conditions spanning input and interpretation.

A key ethical implication follows: if mediation conditions experience and identity, those conditions must be legible and adjustable.

Predictive accounts and the FEP formalise this dependency: perception and action reflect ongoing model–evidence negotiation, and uncertainty minimisation can become maladaptive when systems push users toward premature closure, rigid priors, or habitual over-reliance^[43,44]. The design question is therefore normative, not merely technical: XR and AI reshape the informational and social conditions that guide interpretation, attention, and obligation, thereby shaping the self that is enacted over time. SDT specifies acceptable direction for this influence: assistance should expand autonomy (authorship), build competence (capability, not substitution), and sustain relatedness (trust and belonging)^[45,46].

The interactional requirement of co-determination is what makes such conditioning ethically tractable, and T.A.N. can be read as co-determination principles for an ethical “extended self”:

- **Transparency as Insight:** Because mediation shapes experience, intent, bias, and uncertainty must be visible. Transparency clarifies what is being altered, why, and with what confidence, enabling trust calibration^[33].
- **Adaptivity as Impermanence:** Users develop; support must change with them. Adaptivity tunes (and fades) assistance as goals, context, and competence evolve, avoiding stale assumptions and dependency^[47].
- **Negotiability as Volitional Action:** Users must be able to consent, contest, and override. Without meaningful veto and reversal, systems displace authorship and moral responsibility^[26,45].

Together, this framing shows why T.A.N. is an ethical requirement, not merely a “trust mechanism”: mediated systems must be transparent, adaptive, and negotiable so uncertainty is regulated with users rather than for them.

The ethics of manipulation literature provides a complementary negative argument for these requirements. Philosophical accounts identify three main characterisations of manipulative influence: bypassing rational deliberation, trickery (inducing faulty mental states), and pressure (non-coercive but difficult-to-resist influence)^[48]. Manipulation is widely held to undermine the validity of consent^[49] and to “pervert the way that person reaches decisions, forms preferences, or adopts goals”^[50]. More recent work characterises it as a hidden influence that targets cannot easily become aware of^[51]. These characterisations are directly relevant to XR–AI systems, where perceptual mediation, personalised nudging, and delegated action all create conditions under which influence can become covert, difficult to resist, or substitutive of the user’s own reasoning. T.A.N. can be read as a systematic defence against all three forms. Transparency prevents trickery by making the system’s intent, reasoning, and uncertainty legible, so users cannot be induced into faulty beliefs about what is being influenced or why. Negotiability prevents pressure by ensuring the user always retains a viable exit; consent, override, and revocability eliminate the “awkward and difficult to resist” condition that characterises manipulative pressure^[49]. Adaptivity prevents the subtlest form, bypassing rational deliberation, by ensuring that support engages and progressively strengthens the user’s own deliberative capacities rather than substituting for them; a system that never fades its scaffolding effectively outsources rational deliberation, and over time, such functional outsourcing becomes indistinguishable from bypassing it. Under these conditions, the system’s influence operates not by bypassing or subverting rational deliberation, but by scaffolding it, providing the informational and attentional conditions under which the user can deliberate more effectively while retaining full authorship over the resulting decision. In this reading, co-determination is not merely non-manipulative influence; it is structurally anti-manipulative, because it preserves and strengthens the very deliberative processes that manipulation subverts.

2.3 Extended reality as a perceptual filter: Dependent origination and predictive control

XR, encompassing AR, mixed reality (MR), and VR, is a technological frontier for modulating human perception. Recent work frames XR systems as technologies that can modulate the incoming light field itself rather than merely overlay virtual content, highlighting that XR operates as a perceptual filter acting prior to conscious interpretation^[52]. Beyond addition, XR enables the subtraction or alteration of sensory input through techniques such as Diminished Reality^[53], allowing aspects of the environment to be suppressed or transformed. Together, these capabilities constitute a form of mediated reality in which XR actively filters perceptual evidence rather than passively displaying information. By selectively amplifying, attenuating, or removing stimuli, XR systems shape what users attend to and how they interpret their surroundings. Perceptual filtering should therefore not be treated as a neutral presentation choice, but as an intervention that warrants disclosure of what is being altered and a user-legible rationale for why. This perspective also motivates XR as a systematic testbed for human–AI interaction research. Wienrich and Latoschik^[54] propose an XR–AI continuum and “eXtended AI” arguing that XR can be used to prototype and study prospective AI embodiments and interfaces in controlled, high-fidelity contexts before deployment.

In practice, XR-mediated filtering can be applied in both constructive and protective ways. Constructively, XR can foreground task-relevant cues or reveal otherwise inaccessible information, supporting learning and decision-making^[55,56]. Altered viewpoints and scale transformations can also enable users to reason about spatial or structural relationships beyond ordinary physical constraints^[57–59]. Protectively, XR can suppress distracting or intrusive stimuli, such as visually cluttered backgrounds or advertising content, to reduce attentional fragmentation and cognitive load^[60]. These interventions reconfigure the informational structure of the environment to better align with task demands and user capacity, but they also introduce a distinctive risk: because XR can reshape “ground truth” at the level of perception, the possibility of false beliefs, miscalibrated trust, or socially consequential

omission becomes acute. This motivates treating the co-determination principles (T.A.N.) not as abstract ethical commitments, but as concrete constraints for responsible perceptual regulation.

- **Transparency in Perception:** The system must disclose how it is filtering reality. If an XR system suppresses visual noise (e.g., removing ads or clutter^[60]), users must be aware that information is being hidden and why. Perceptual transparency prevents users from mistaking a curated evidential stream for objective reality.
- **Adaptivity in Scaffolding:** Perceptual enhancements, such as highlighting task-relevant cues^[55], should not become fixed or miscalibrated supports. True adaptivity implies that as a user learns to notice patterns (increasing perceptual competence), highlights can fade or re-target, transferring predictive load back to the user while preserving support when conditions change.
- **Negotiability of Reality:** Users must have the power to define and revise their perceptual boundaries. Whether it is a therapeutic application modulating anxiety triggers or exposure in Post-Traumatic Stress Disorder (PTSD) treatment^[61], or a productivity tool filtering distractions, users should be able to inspect, override, and revert filtering on demand, including simple “show me what was removed” controls.

First, XR can function as a perceptual enhancer to reduce surprise by providing timely, task-relevant cues that make situations more predictable. If perception is, as Andy Clark suggests, a kind of “controlled hallucination” constrained by sensory feedback, then XR can be understood as an externalised intervention in the evidence that stabilises perceptual inference^[62]. By enhancing signal quality or suppressing noise, XR systems shape the feedback that constrains expectations, making environments more predictable and cognitively manageable. Examples include AR navigation overlays that reduce wayfinding ambiguity^[55], military helmet displays that stabilise situational awareness in fast-changing environments^[63], and surgical AR systems that integrate imaging data directly into the operative field^[64]. However, cueing can also become static or miscalibrated if it is not designed to adapt as competence develops. Adaptivity, therefore, implies scaffolding that can be intensified, faded, or re-targeted as user skill and context change, transferring predictive load back to the user where appropriate while preserving support when conditions change.

Conversely, XR can be used as a perceptual filter to minimise surprise or stress by attenuating extraneous or harmful inputs. The Ad-Blocked Reality prototype, for instance, digitally obscured billboard advertisements and reported increased focus and subjective calm while raising concerns about the removal of potentially relevant information^[60]. More broadly, XR interventions have been characterised as perceptual manipulations that can extend beyond immediate action to influence longer-term cognitive processes, including memory and self-perception^[65]. In therapeutic contexts, XR-based filtering has also been proposed as a means of attenuating anxiety triggers or modulating exposure in Post-Traumatic Stress Disorder PTSD treatment, enabling gradual recalibration of expectations under controlled conditions^[61]. This illustrates a central trade-off: while filtering can enhance immediate well-being and a sense of control, excessive or opaque filtering risks distorting the user’s world model and narrowing attention in ways analogous to perceptual echo chambers. Negotiability is therefore not optional. Users should be able to inspect, override, and revert filtering on demand, including simple “show me what was removed” controls, especially where safety, fairness, or civic relevance are at stake.

These ethical requirements become even clearer in fully synthetic VR. By presenting a largely constructed sensorium, VR allows systematic manipulation of the relationship between expectation and sensation, making the role of prior beliefs in shaping experience explicit. Classic embodiment illusions, such as virtual limb or full-body ownership, arise when visual and sensorimotor contingencies align with the brain’s expectations, leading users to experience virtual bodies as their own^[66]. The resulting sense of presence, the feeling of “being there”, can be understood as successful perceptual inference that the virtual world is sufficiently real to act within^[67,68]. Importantly, compelling experience depends not only on rendering fidelity but also on behavioural and narrative coherence, because incoherent cues can collapse plausibility even in highly immersive systems^[69]. This implies a transparency obligation that goes beyond “what was rendered”: systems should help users distinguish evidential cues from narrative framing, and support stepping out of persuasive framing when desired.

Empirical studies show that controlled manipulation of perceptual evidence can yield lasting changes in internal models. Rastelli *et al.*^[70] found that VR-induced visual hallucinations increased cognitive flexibility by loosening entrenched perceptual priors, while another study^[71] showed that a placebo VR intervention that visually exaggerated physical capability produced sustained real-world motor gains by elevating expectations about bodily ability. In predictive-processing terms, both effects reflect prior updating through structured prediction errors that can persist beyond the session. This strengthens the case for consensual, goal-aligned interventions that build competence rather than dependency, especially when altered expectations may generalise to everyday life.

Taken together, these examples show XR acting as both a perceptual enhancer and filter, enabling direct intervention in the inferential processes that generate perception by adding, removing, or restructuring sensory evidence. Such mediation parallels operant conditioning, where stimulus presence or removal guides learning and behaviour^[72], a mechanism already leveraged in XR design to intentionally engage or disengage users^[73]. To avoid drifting from support into behavioural control, reinforcement intent should be explicit, auditable and user-configurable in high-stakes contexts.

XR thereby makes tangible the dependent-origination insight that experience is conditioned, while predictive processing and FEP explain how altered evidence reshapes inference over time^[42,44]. Recent AR work also operationalises SDT directly by testing how adaptive assistance shifts perceived autonomy: in AR-assisted construction assembly, low-agency control reduced workload but also reduced perceived autonomy, highlighting the agency trade-off that Self++ is designed to manage^[74]. Co-determination then specifies the interactional obligation for XR perceptual filtering: because XR can alter the evidential conditions of experience, users should be able to recognise what has been amplified, attenuated, or removed, why that regulation is occurring, and how to inspect, revise, or reverse it. In this way, perceptual support can remain aligned with autonomy, competence, and relatedness rather than becoming a covert form of behavioural control.

2.4 Human–AI interaction and teaming in XR

The preceding sections framed XR as a mechanism for intervening on the evidential conditions of experience (through perceptual filtering), and situated Self++ within a broader philosophical view in which experience and identity are enacted through relational conditions. When we move from perception to action, the locus of risk and opportunity shifts: the AI is no longer merely shaping what is seen or attended to, but increasingly participates in goal selection, planning, and execution. This transition places Self++ within Human–Agent Teaming (HAT) or Human–AI Teams (HATs), which studies how humans and autonomous systems coordinate to achieve shared goals^[1,75,76]. In XR and metaverse-like settings, teaming is not only informational but embodied and situated: coordination unfolds through shared spatial context, sensorimotor coupling, and the ongoing regulation of cognitive load and uncertainty.

A central obstacle for effective teaming is the user’s difficulty in forming accurate mental models of an agent’s state and reasoning, a challenge Norman characterises as the “gulf of evaluation”^[77]. In XR, this gulf can widen: immersive presentation may increase perceived immediacy and credibility, while the agent’s internal uncertainty, constraints, and operating assumptions remain hidden. This is precisely where the interactional stance of co-determination becomes necessary. Rather than assuming fixed tool use or unilateral automation, co-determination treats human and agent as joint participants in a coupled system, requiring that the agent’s intent, boundaries, and uncertainty be legible enough for the user to retain volitional control. Users also carry expectations about what a “good” AI teammate should be^[76] shows that people often expect AI partners to behave with human-like reliability, cooperativeness, and contextual sensitivity, and mismatches between these expectations and actual system behaviour can undermine trust and coordination. These expectation dynamics strengthen the case for co-determination as a stabilising baseline: the system must help users calibrate what the agent can and cannot do, rather than letting anthropomorphic assumptions silently drive reliance. Building on this trajectory, cognitive externalisation is now evolving into adaptive agent teammates, where calibrated trust shapes effective interaction and HAT outcomes^[78].

In this practical domain of teaming, the co-determination principles (T.A.N.) must be implemented as specific interaction mechanisms rather than treated as abstract ethical principles:

- **Transparency for bridging the gulf of evaluation:** The agent should make its internal state legible enough for users to form accurate mental models, including what it is optimising for, what it believes, and where uncertainty or constraints apply. This reduces evaluation gaps and supports trust calibration^[19,24,77].
- **Adaptivity for dynamic allocation of initiative:** Effective teammates do not behave identically regardless of context. Agents should adjust initiative, timing, and level of autonomy as user confidence, workload, and task conditions change, supporting decision outcomes without overwhelming or bypassing the user^[56].
- **Negotiability for consensual delegation and recovery:** As agents become more capable, the risk of automation bias and loss of control increases. Users should be able to consent to actions, revise autonomy levels (e.g., “help me do this” versus “do this for me”), and override or undo decisions, preserving authorship and accountability^[79].

General human–AI interaction guidance reinforces these requirements. Established guidelines emphasise making clear why the system acted, supporting efficient correction, and enabling undo and refinement^[47]. In XR, where the system can shape both evidence and action, such principles are not cosmetic: they protect autonomy and competence by reducing surprise, supporting trust calibration, and preventing opaque shifts in control. Recent empirical work on team dynamics in human–AI collaboration further emphasises that teaming outcomes depend on interaction quality, affecting confidence, satisfaction, and accountability^[24]. From a Self++ perspective, these are not merely usability metrics; they indicate whether an agent supports or frustrates SDT needs, and whether the coupled system converges towards stable, low-surprise coordination.

This interactional framing is consistent with XR-specific work on explanation and intelligibility. The XAIR framework^[80] for explainable AI in AR argues that systems should generate explanations with AI outcomes and keep them accessible to support user agency, while using manual, user-triggered delivery as the default due to limited cognitive capacity in AR. XAIR further recommends that automatic, just-in-time explanations be reserved for constrained cases (e.g., surprise or confusion, unfamiliar outcomes, or model uncertainty) and only when the user has enough capacity to attend to them. Beyond timing, XAIR emphasises end-user

configuration and a longer-term user-in-the-loop co-learning process, where systems adapt to users while users' understanding and AI literacy evolve. In HAT terms, these design commitments instantiate the co-determination principles (T.A.N.): explanation access and state-legibility as *Transparency*, timing and initiative control as *Adaptivity*, and user-trigger, configuration, and reversibility as *Negotiability*.

Recent XR-specific HAT work further illustrates how embodied context changes the nature of coordination. Zhang *et al.*'s "Virtual Triplets" framework^[18] analyses dynamics between the human, the virtual agent, and the physical task across synchronous and asynchronous settings. Successful assistance requires sensitivity to physical constraints, task progress, and translation between digital instruction and physical execution, aligning with the Competence overlay of Self++; the agent's role is not to replace skill, but to scaffold effective action^[34]. XR training research demonstrates this scaffolding role in practice. HAT Swapping^[19] explores how virtual agents can act as stand-ins for absent human instructors, enabling guidance and feedback to persist across time and personnel while preserving the structure of collaborative training. AVAGENT^[81] similarly shows how AI-powered virtual avatars can bridge asynchronous communication by capturing, transforming, and re-presenting human intent and context across time, extending HAT beyond real-time copresence into persistent coordination in XR. Together, these systems highlight both the promise and responsibility of XR agents: they can reduce uncertainty and support skill acquisition, but only if guidance remains transparent, appropriately timed, and adjustable to the learner's evolving competence.

As agents become more capable, the design challenge intensifies. Multimodal foundation models enable systems that can perceive and act across vision, audio, language, and contextual signals, supporting increasingly high-level delegation^[82]. However, increased capability increases the risk of misalignment and opacity, especially when the user cannot inspect the agent's evolving beliefs or intentions. Work on transparency for modern AI systems emphasises interactive scrutability, user education, and attention to socio-cultural context as prerequisites for trustworthy deployment^[83]. In Self++ terms, this is a direct extension of co-determination: an agent that can act must remain accountable through transparency and negotiability, allowing users to query intent, adjust autonomy, and recover from errors. Approaches that explicitly structure agent reasoning, such as BDI-style models^[84], can support this by making beliefs, desires, and intentions more inspectable, thereby narrowing the gulf of evaluation and reducing the likelihood that delegation undermines user agency.

These issues are not unique to XR, and lessons from human-AI co-creation generalise. Studies of collaborative writing with language models highlight recurring problems of trust calibration, user control, and authorship, even in ostensibly low-stakes tasks^[85]. Recent work on agency in large language model (LLM)-infused tools similarly suggests that preserving authorship depends on making suggestions legible and easy to veto, so that assistance remains subordinate to the user's intent rather than silently steering outcomes^[86]. These findings map naturally onto the Autonomy overlay of Self++ and provide concrete interaction criteria for co-determination in XR: the agent should behave as a co-pilot whose contributions are inspectable, revisable, and aligned with the user's chosen level of control. Shneiderman's Human-Centred AI perspective reinforces this stance, arguing for systems that combine powerful automation with meaningful human control and accountability rather than pursuing full autonomy as an end in itself^[87]. Yang *et al.*^[88] likewise note that human-AI interaction is uniquely difficult to design due to unpredictability, feedback gaps, and mental-model mismatch, highlighting the need for iterative design processes to tame agent behaviour in practice. In XR, where the system can shape both action and perception, these difficulties are amplified, strengthening the case for co-determined interaction as a baseline expectation.

Finally, the social dimension of teaming is essential, particularly for the Relatedness overlay of Self++. Triadic human-agent dynamics show that agents can mediate human-human collaboration, influencing how people coordinate and communicate with one another^[79]. Embodied virtual agents can elicit prosocial responses by expressing social cues and emotional responsiveness, suggesting a pathway for agents to scaffold social functioning rather than merely simulate companionship^[89]. More generally, empirical work in AR suggests that visual embodiment and social behaviours shape how intelligent assistants are perceived, influencing credibility, social presence, and users' willingness to engage with or rely on an agent^[90]. At the same time, work on "generative agents" shows how AI characters with memories and routines can produce richly believable social dynamics in simulated worlds^[91]. Such agents could enrich relatedness in XR by enabling meaningful social rehearsal, community participation, or culturally situated narratives; yet they also sharpen the risks of uncanny, unpredictable, or normatively misaligned behaviour that can destabilise user trust and undermine a sense of agency. From the philosophical framing established earlier, this is not peripheral: if experience and identity are enacted through relational conditions, then agent-mediated social worlds will shape the kinds of selves that become habitual. Co-determination, therefore, extends beyond individual task control to relational accountability: agents that participate in social coordination should be designed to strengthen users' connection to real communities and shared norms, rather than displacing human relationships through frictionless substitutes^[21].

Taken together, HAT in XR offers the interactional mechanisms through which Self++ can be realised across the three overlays: competence, autonomy, and relatedness. XR can reorganise sensory evidence and reduce uncertainty, but as AI shifts from filter to collaborator, the conditions for healthy regulation of uncertainty become fundamentally interactional. Co-determination provides the bridge from the cognitive and philosophical foundations to concrete HAT practice: by prioritising the co-determination principles

(T.A.N.), XR agents can scaffold skill, preserve volitional control, and strengthen social embeddedness, rather than causing relational displacement.

3. The Self++ Architecture: Three Overlays of Augmented Agency

Self++ organises human–AI coupling into three concurrently activatable overlays (Self, Self+, Self++), forming an architecture of augmented agency (Figure 1). Each overlay targets a different temporal and functional scale of free-energy minimisation, corresponding to nested timescales of adaptation and echoing “nested learning” in AI^[92]. The naming (Self, Self+, Self++) does not imply separate selves, but an expanding *scope* of agency support: from here-and-now action to deliberation and policy formation, to social embeddedness.

Importantly, Self++ does *not* assume a strict pipeline in which Overlay 1 must finish before Overlay 2 or Overlay 3 begins. In realistic settings (training, teamwork, community participation), competence-building, autonomy exercise, and relatedness-support often co-occur, and initiative shifts fluidly between human and system^[93,94]. We therefore treat the three overlays as concurrently activatable modes of a coupled human–AI system, with different emphases depending on context and risk.

A clarification is important here: the temporal-horizon labels, short-, intermediate-, and long-term horizons (Table 1), denote where each overlay’s design commitments are *primarily* anchored, not where they are exclusively confined. A Tutor interaction may unfold in seconds to minutes per episode, while a tutoring relationship persists for months; what anchors the Tutor role at the sensorimotor timescale is that its key design variables, such as cue timing, step gating, and attention regulation, are specified and evaluated at that temporal grain. Conversely, a Social Facilitator primarily operates at the relational timescale while still needing to respond in real time to conversational dynamics. The overlay labels, therefore, indicate the *primary design horizon* for each set of role patterns, not a boundary on when they may be active.

Table 1. Self++ role patterns across overlays (concurrently activatable overlays) with example XR-AI behaviours and co-determination principles (T.A.N.).

Lvl	Role	Role objective	Example XR-AI behaviours	Transparent	Adaptive	Negotiable
Overlay 1 (Self): Competence support (short-horizon)						
R1	Tutor	Reduce novice uncertainty; establish safe learnable corridor	Anchored arrows and ghosted exemplars with step gating; clutter suppression; completion detection with attention-aware pacing and corrective feedback	Cue provenance; disclose suppression; show limits	Fade prompts; retarget errors; adjust pacing	Pause/skip; show all vs minimal; override highlights
R2	Skill Builder	Calibrate + generalise; variability with feedback, not scripting	Ghost tracks and shadow end-states with partial hints; performance analytics with adaptive hinting and controlled variability	Explain feedback basis; show comparison model	Increase task variability; withhold hints; change modality	User-set difficulty; toggle ghosts; consent for perturbations
R3	Coach	Robustness under stress; self-correction; prevent brittle mastery	Fault injection and overlay removal with altered timing; safety/quality monitoring; targeted debrief with fall-back to R1/R2	Disclose perturbation intent; disclose role/agency shifts	Adjust challenge intensity; adapt thresholds; taper monitoring	Opt-in for stress tests; emergency stop; hand-off confirmation
Overlay 2 (Self+): Autonomy support (intermediate-horizon)						
R4	Choice Architect	Shape decision context (salience) while preserving authorship	Lightweight cueing with route salience (alternatives remain selectable); multi-criteria filtering; attention weighting with trade-off previews	Mark nudges; link to goals; label optimised criteria	Update weights; fade as user internalises; reduce during load	Opt-out slider; consent for high-stakes; unmudged view
R5	Advisor	Externalise deliberation; make	Interactive dashboards with side-by-side futures and	Expose sources; distinguish	Tune depth; switch modality; calibrate to	Editabale goals; ask-for-alt; decline

		counterfactuals inspectable	uncertainty bands; value elicitation; model explanation with alternatives and effect highlights	evidence vs framing; show unknowns	time pressure	reasoning; override defaults
R6	Agentic Worker	Delegated execution under user policy; proposal-approval loop	Plan and execute with XR review checkpoints; plan trace with progress visibility; step confirmation with safe interrupts and rollback	Show intent/plan; audit trail; capability limits; risk disclosure	Adjust frequency by stakes; learn checkpoints; degrade gracefully	Explicit delegation; revoke anytime; re- scope; adjustable autonomy
Overlay 3 (Self++): Relatedness & purpose (long-horizon)						
R7	Contextual Interpreter	Legibility of identity/norms + impacts; reduce social surprise	Human vs AI labels and role badges; provenance overlays with impact cards; norm reminders; plural framing for contested topics	Radical disclosure of agent identity, show provenance	Context density tuned to attention; adapt to culture/values	Controls for context appearance; sensitivity sliders; opt-out
R8	Social Facilitator	Improve coordination + repair; increase human-human connection	Shared gaze and participation balance visualisation; micro- clarifications; breakdown detection with viewpoint summaries and perspective- taking prompts	Disclose sensing granularity; explain prompts + thresholds	Do-nothing mode when thriving; calibrate to group norms	Collective opt-in; privacy-by-role; group-negotiable
R9	Purpose Amplifier	Long-horizon value coherence; steer away from disavowed futures	Value-facing simulations with nudges-in-narrative and framing controls; periodic reflections; contestable inferences with governance hooks	Reason + framing legibility; evidence vs narrative separation	Internalisation- focused fading; calibrate identity strength	Contestability; escalation requires opt-in; collective pathways

XR: extended reality; AI: artificial intelligence.

Overlay 1 (Self): Competence at the sensorimotor timescale (short-horizons). This overlay augments perception and skill, reducing immediate prediction errors in action execution^[33].

Mechanistic coupling (SDT-FEP): Competence ↔ minimisation of sensorimotor prediction error. Competence is the subjective experience of a highprecision internal model effectively governing action. When the AI scaffolds skill (e.g., highlighting a target), it reduces the gap between predicted and actual sensory feedback, validating the user's model of agency^[34,95].

Overlay 2 (Self+): Autonomy at the deliberative and situational timescale (intermediate-horizon). This overlay augments cognition and decision-making, helping users navigate complex choices and intermediate goals by reducing strategic uncertainty^[32].

Mechanistic coupling (SDT-FEP): Autonomy ↔ preservation of high-level priors (policy selection). Autonomy reflects the ability to self-endorse actions. In FEP terms, this equates to selecting policies aligned with deep, top-down priors (values/goals), rather than being driven by bottom-up salience or external coercion^[32]. AI support here reduces “decision entropy” while protecting the user's generative model from being overwritten by the system^[87].

Overlay 3 (Self++): Relatedness at the developmental and existential timescale (long-horizon). This overlay augments social connection and purpose, steering long-term trajectories and relationships by aligning actions with enduring values and shared social models^[96].

Mechanistic coupling (SDT-FEP): Relatedness ↔ alignment of shared generative models. Relatedness arises from synchronisation of internal models between agents: social connection enables partial offloading of uncertainty onto the group. AI support here minimises “social surprise” (misinterpretation of others) and helps the user remain embedded in a shared communicative web^[42,97].

A methodological note on these couplings: SDT and FEP operate at different levels of description; SDT is a motivational theory

grounded in decades of experimental psychology, while FEP is a formal account of biological self-organisation rooted in variational inference. The correspondences proposed above (competence ↔ sensorimotor prediction-error minimisation; autonomy ↔ high-level prior preservation; relatedness ↔ shared generative-model alignment) are bridging hypotheses that constitute part of this paper's theoretical contribution, not established equivalences. We propose them because they generate productive design commitments (the overlays and T.A.N. constraints) and because they yield testable predictions (Table 2). However, the mapping is neither unique nor exhaustive: alternative bridging constructs are possible, and empirical work may reveal that the coupling is tighter for some needs than others or that additional mediating constructs are required. We therefore treat these correspondences as working hypotheses to be refined through the evaluation programme outlined in Section 7 and the future work discussed in Section 9.5.

Table 2. Self++ propositions (P1–P8) with brief evaluation checks. Use as a lightweight audit: map system features to Self++ role patterns, then verify T.A.N at the required strength and test transitions and drift under realistic overlap.

P	Proposition (what must be true)	Evaluation checks (what to test/measure)
P1	Concurrency: Overlays act concurrently (not a pipeline) and can interfere.	Test overlap interference and recovery: (i) run Overlay 1 guidance while Overlay 2 deliberation UI is present (e.g., motor task + counter-factual dashboard) and measure errors/time-on-task; (ii) measure reclaim-time (time to pause/override after an AI-led phase) and success rate of taking back control.
P2	Timescale Alignment: SDT needs map to uncertainty targets across temporal scales.	Evaluate on the right horizon: Overlay 1 with immediate sensorimotor metrics (errors, collisions, smoothness); Overlay 2 with decision quality and goal-alignment/endorsement (regret, confidence, stated-goal match over days); Overlay 3 with longitudinal drift indicators (relationship repair, wellbeing, dependence, value-consistency) over weeks/months, not only short task scores.
P3	Inspectability: Legitimate augmentation requires an inspectable, contestable AI voice.	Probe legibility and ownership: users can state what was influenced (evidence, salience, delegation), why, and how to reverse the present intervention. Behavioural test: can users successfully access alternatives, inspect reasons, and undo or suspend the current support?
P4	T.A.N. Scaling: Co-determination strength must scale with scope and initiative.	Audit proportional safeguards: higher-scope interventions in Overlay 3 must provide stronger provenance and incentive disclosure, clearer consent boundaries, broader reversibility, and more complete audit trails than lower-scope Overlay 1 support. Test whether safeguard strength increases appropriately with intervention scope and initiative.
P5	Transition Legibility: Shifts in agency between role patterns must be perceptible and reversible.	Test hand-offs and escalation/de-escalation: users must correctly identify when agency has shifted, who is acting, and under what authority. Measure transition awareness, misattribution rates, and recovery after failed or unwanted hand-offs.
P6	Endorsement over Compliance: autonomy support preserves authorship over revision, not mere compliance.	Check internalisation, not just performance: users endorse outcomes as their choice and can explain “because...” in terms of their goals/values. Compare nudged vs unnudged conditions: if outcomes improve but endorsement drops or users cannot justify choices, autonomy support failed.
P7	Collective Negotiability: Relatedness support requires shared-model alignment and group negotiability.	Verify group legitimacy: collective opt-in for sensing/visualisations; privacy-by-role defaults; and opt-out without social penalty (no status loss, no exclusion cues). Test whether participants can contest aggregation rules/thresholds (e.g., participation metrics) and still collaborate smoothly.
P8	Governance Contestability: Long-horizon alignment is socio-technical and requires contestation pathways.	Audit contestability of action and framing: users and affected groups can challenge not only recommendations but also optimisation targets, interpretive categories, escalation criteria, and institutional defaults. Verify pathways for review, appeal, and collective contestation where communities are affected.

How to use (self-contained): (1) Map features to Self++ role patterns (R1–R9) across the three concurrently activatable overlays (Table 1); (2) For each claimed role pattern, verify co-determination (T.A.N.) commitments at the required strength (reasons/provenance/incentives; fading/calibration; override/contestability); (3) Evaluate transitions and long-horizon drift under realistic concurrent operation, not only steady-state task performance. Evidential status: Section 7.2 maps each proposition to its current empirical support (direct, indirect, or open hypothesis) and identifies evaluation priorities. SDT: Self-Determination Theory.

Within each overlay, Self++ specifies three role patterns (R1–R9 in total), each realised as an AI role that supports the user under co-determination. R1–R3 correspond to *Tutor*, *Skill Builder*, *Coach* (Overlay 1); R4–R6 to *Choice Architect*, *Advisor*, *Agentic Worker* (Overlay 2); and R7–R9 to *Contextual Interpreter*, *Social Facilitator*, *Purpose Amplifier* (Overlay 3). Together, these nine role patterns form the core of the Self++ framework (summarised in Table 1). The decomposition into named role patterns, rather than a single continuously adaptive agent or an undifferentiated spectrum of support, serves three design functions aligned with co-determination. First, discrete roles make the system’s current supportive intent legible to the user: knowing that the system is acting as a Coach (testing robustness) rather than a Tutor (guiding a novice) allows the user to calibrate expectations and interpret feedback correctly, directly supporting Transparency. Second, bounded role definitions give designers and evaluators discrete interaction contracts against which to assess T.A.N. compliance and test the propositions in Table 2; a diffuse, unlabelled adaptation would be harder to audit or compare across implementations. Third, named roles provide natural anchors for Negotiability: users can request transitions (“stop advising, just execute”), decline specific modes (“no stress tests today”), or query the system’s current stance (“why are you coaching rather than building?”), interactions that would be less intuitive if support varied along an unmarked continuum. We note that this decomposition is functional, not architectural: whether the nine roles are implemented as a single foundation model selecting behavioural policies, as composable modules, or as a mixture of both is an engineering decision left to implementation. What matters for Self++ is that the user-facing interaction contract of each role remains distinct and legible regardless of the underlying system design. While functional, roles use everyday personifications (such as a tutor or a coach) to clarify supportive intent rather than authority.

The overlays should also not be understood as merely coexisting in parallel. In practice, they actively shape one another. Gaining clarity about what one values (Overlay 3) can reveal new skills worth developing (Overlay 1) and reframe choices about how to pursue them (Overlay 2). Conversely, building new competence (Overlay 1) can expand what options feel available in deliberation (Overlay 2) and, over time, reshape identity, commitment, and purpose (Overlay 3). This recursive dynamic of doing, choosing, and becoming means that the self interacting with Self++ at month six is not identical to the self that began at month one. Self++ accommodates this by treating overlays as concurrently activatable and mutually permeable: outputs from one overlay, such as a refined value commitment in Purpose Amplifier (R9), can become updated inputs to another, such as new learning goals for Tutor (R1). An important direction for future work is to investigate this generative cycling empirically, tracing how interventions at one overlay propagate through the others over longitudinal timescales.

Crucially, role patterns act as adaptive scaffolds: as competence, context, and risk change, the system transitions between role patterns or fades support to prevent over-reliance and to preserve human autonomy and relationships^[98]. To keep augmentation legitimate rather than covert control, Self++ applies the co-determination principles (T.A.N.) across all overlays:

- **Transparency:** Sufficient information for accurate mental models of intent, limits, incentives, and uncertainty.
- **Adaptivity:** Support tuned over time as competence and context evolve (including fading).
- **Negotiability:** Volition preserved via consent, override, and adjustable autonomy.

T.A.N. requirements strengthen with scope and initiative: higher-overlay role patterns (especially those touching identity, relationships, or long-horizon behaviour) demand stronger transparency and negotiability as safeguards^[93,94].

4. Overlay 1 – Foundational Augmentation of the Self (Competence Support)

Overlay 1 targets competence at the sensorimotor timescale: helping users perceive and act reliably in an enriched environment, while keeping early errors and overload low enough for learning to take hold. Self++ does not treat this as a prerequisite pipeline stage. Competence support often runs in parallel with autonomy and relatedness supports (for example, training in teams), but Overlay 1 remains the point where the system most directly shapes perceptual evidence and action feedback.

Mechanistically, Overlay 1 reduces sensorimotor prediction error so users experience effectance and learnable control: attention is guided, actions are constrained into safe steps, and feedback tightens the link between intention and outcome. In SDT terms, this sustains competence by enabling early, attributable successes; in FEP terms, it increases the precision of action–outcome mappings and reduces surprise during control^[16,32]. We define three role patterns that mirror established progressions in skill acquisition from novice to proficient performance: *Tutor* (R1), *Skill Builder* (R2), and *Coach* (R3)^[99]. As with higher overlays, these role patterns are co-determined and scaffolded in line with the co-determination principles (T.A.N.), where support should be *Transparent* (users can tell what is being guided and why), *Adaptive* (fading as competence stabilises), and *Negotiable* (users can slow, pause, or override guidance), so assistance accelerates learning without converting into dependency.

4.1 Role pattern R1: Guided familiarisation (AI as Tutor)

At the outset of a new task or environment, novices face high uncertainty because relevant cues, action boundaries, and error consequences are not yet well-modelled. In the Tutor role, the AI adopts a proactive stance that structures the experience into a learnable corridor: it highlights what matters, suppresses what is distracting, and sequences actions so that each step is achievable

before the next is introduced. This is classic scaffolding in the Zone of Proximal Development^[34], but implemented through in-situ perceptual guidance rather than detached instructions.

In XR, this guidance can be spatial and embodied: key objects or regions can be highlighted, next actions can be indicated with anchored arrows^[100] or ghosted exemplars^[101], and irrelevant elements can be visually deemphasised to reduce split attention. A practical pattern is step gating: the system reveals only the next required sub-action and advances when completion is detected, which keeps working memory demands bounded. Adaptive AR tutoring systems have operationalised this idea by monitoring tutorial-following status and adjusting the amount and form of guidance in real time^[102]. When attention lapses, a Tutor can also regulate pacing through attention-aware playback (for example, pausing or slowing guidance when gaze or location cues indicate the user has fallen out of sync), helping the user recover without compounding errors.

Technically, the Tutor role overlaps with intelligent tutoring systems that use cognitive models to interpret learner actions and deliver context-sensitive feedback (for example, model tracing and related methods in cognitive tutors)^[103]. The key difference in XR is that feedback can be embedded directly into the perceptual field, allowing guidance to be shown where and when it is needed rather than translated into verbal rules.

Empirical evidence supports the value of structured, in-situ guidance during early skill acquisition. In assembly-like tasks, AR instructions have been shown to reduce errors and improve performance relative to conventional instruction formats in controlled comparisons^[104]. At the same time, the broader literature cautions that AR can either reduce or increase cognitive load depending on design choices, which strengthens the case for tightly scoped, well-timed guidance at R1^[105].

Finally, the Tutor role pattern is designed as deliberately temporary for users whose capacity and goals support progression: as soon as the user demonstrates stable performance on a step, guidance should begin to fade (fewer cues, larger action windows, more self-explanation) and the system should transition towards the Skill Builder role pattern. This aligns with evidence from instructional design that gradually reducing worked guidance supports the learner's shift from example-following to independent problem solving^[106,107]. In summary, R1 establishes a safe learning corridor with immediate, situated success signals, while explicitly preparing the conditions for the scaffold to be removed.

4.2 Role pattern R2: Scaffolded practice (AI as skill builder)

Once the user can complete the basic sequence under guided familiarisation, the AI shifts into the Skill Builder role pattern that prioritises practice, calibration, and generalisation. The support envelope deliberately widens: the system provides partial cues and performance feedback but stops prescribing every micro-action. The intent is to refine the user's sensorimotor predictions while avoiding the brittleness that comes from rehearsing a single, fixed script. Motor-learning theory predicts that variability and appropriately structured interference during practice can improve transfer and retention, even if acquisition feels harder^[108,109].

A hallmark of R2 is augmented feedback that keeps “what good looks like” visible while leaving execution to the user. Two common XR patterns are Ghost Tracks, which overlay time-aligned expert motion for in-situ trajectory and timing matching^[17,110–113], and Shadow Workspaces, which anchor a target end-state silhouette (“shadow of success”) to support precise pose, placement, or orientation^[17,105,114,115]. Together, they externalise comparison and reduce cognitive load during repeated practice while preserving active control.

Although these cues are most natural for 3D sensorimotor tasks, the underlying principle generalises: externalised reference structure reduces internal memory and computation by making intermediate steps, trajectories, or goal states inspectable^[8]. In non-spatial skills, this parallels worked examples and step-by-step solution traces that support novice learning and later independent problem solving^[116,117]. Seen this way, Ghost and Shadow patterns are XR instantiations of a broader apprenticeship logic of modelling, scaffolding, and fading across embodied and cognitive procedures^[103,118].

Crucially, R2 also introduces controlled challenge. Rather than maximising ease, the system should keep the task in a learnable difficulty band by gradually withholding hints, expanding acceptable action ranges, and introducing mild perturbations (for example, small changes in order, timing constraints, or plausible micro-faults) so the user learns to adapt rather than imitate. This “challenge just beyond current mastery” is consistent with the challenge-skill balance emphasised in flow-oriented accounts of engagement and growth^[119]. It also parallels curriculum ideas from machine learning, where a teacher proposes goals that are increasingly difficult but achievable, as in AMIGO^[120]. In Self++, the Skill Builder role pattern therefore balances error reduction with productive difficulty: enough structure to prevent unproductive surprise, enough freedom and variability to build robust competence. By the end of R2, the user should rely on Ghost and Shadow cues primarily for fine-tuning, while completing substantial portions of the task without explicit step-by-step prompting.

However, Self++ does not assume that all users will or should progress beyond R2. For individuals whose capacities, contexts, or preferences make sustained scaffolding the appropriate endpoint, including many users with disabilities who experience assistive technologies as extensions of self rather than temporary supports, remaining at R2 long term is a valid, competence-affirming outcome. What matters is whether the level of support is aligned with the user's endorsed goals and current capacity, not whether it

matches an externally imposed trajectory toward independence.

4.3 Role pattern R3: Mastery and resilience (AI as coach)

Once the user is reliably proficient in routine conditions, the AI transitions to the Coach role pattern focused on robustness, adaptability, and self-correction. Guidance recedes: instead of persistent highlights or continuous overlays, the Coach monitors performance and introduces controlled perturbations to test whether the skill generalises beyond rehearsed cases. This deliberate use of “desirable difficulties” supports more durable, flexible learning than perfectly predictable practice^[121,122] and matches accounts of expertise that emphasise deliberate, feedback-rich refinement over time^[123].

In practice, the Coach varies scenarios, injects plausible faults, and occasionally withholds support (for example, removing an overlay or altering timing constraints) to expose brittle assumptions and reveal blind spots. It intervenes only when performance drops below a safety or quality threshold, preventing the consolidation of poor habits while keeping the user responsible for recovery and strategy. After each episode, the Coach provides a brief debrief and, if needed, temporarily reverts to Tutor or Skill Builder to remediate a specific sub-skill. In Self++ terms, R3 consolidates competence by reducing “surprise under stress”: the user learns not only to execute correctly, but to remain stable when conditions deviate from expectation^[122].

R3 also manages role transitions in team settings, so the user retains a coherent model of who is doing what. Abrupt hand-offs, silent autonomy shifts, or ambiguous identities can trigger mode confusion and automation surprise, especially in off-nominal situations^[124,125]. Self++ operationalises this as HAT Swapping^[19]: a user-legible protocol for transferring a functional role between human and AI teammates (and back again), preserving continuity cues where helpful while explicitly disclosing changes in agency and identity so trust remains calibrated^[126]. When the coach function is embodied (for example, via an avatar), continuity cues (voice, interaction style, interface layout) can smooth transitions, but disclosure of boundaries and identity remains essential, in line with the co-determination principles (T.A.N.): *Transparency* about what changed and why; *Adaptivity* in how much coaching is offered; and *Negotiability* to prevent over- or under-reliance^[127,128].

By the end of R3, the user should display functional mastery: resilient performance across varied conditions, recovery from errors without constant prompting, and correctly calibrated trust in the coach as a safety net rather than a crutch.

5. Overlay 2 – Cognitive and Strategic Augmentation (Autonomy Support)

Overlay 2 shifts emphasis from executing skills to forming and revising policies: choosing goals, weighing trade-offs, and allocating attention and effort over time. Self++ does not treat the three overlays as a strict pipeline. Autonomy support often appears during competence building: even in training, learners must make meaningful choices (what to try next, when to speed up, whether to accept risk, when to request help) in order to demonstrate genuine competence. Accordingly, Overlay 2 can run concurrently with Overlay 1: the system may coach sensorimotor execution while also shaping the user’s decision context so choices remain aligned with the user’s own values and intentions.

This concurrent view matches mixed-initiative and adjustable-autonomy systems, where initiative and control shift fluidly between human and agent depending on task demands, user state, and risk, rather than advancing through fixed stages^[93,94,129]. Mechanistically, Overlay 2 targets autonomy as policy selection: in SDT, autonomy is experienced as self-endorsed action^[32]; in FEP terms, this corresponds to protecting high-level priors (values and goals) while using prediction to reduce uncertainty about consequences^[16]. In Self++ terms, Overlay 2 is co-determination expressed at the cognitive timescale: a second voice that helps the user reflect, anticipate outcomes, and surface trade-offs, but does not smuggle in new goals or override the user’s higher-order commitments. This caution is reinforced by evidence that synthetic persuasion evaluations can diverge from human outcomes^[20,130].

A key autonomy risk in modern ecosystems is that choice environments are routinely shaped by opaque recommendation logic, engagement optimisation, and dark-pattern design, steering behaviour while eroding the user’s sense of authorship^[131,132]. Self++, therefore, requires decision support to remain co-determined: (i) legible enough for users to judge how the system is weighing attention and effort, (ii) responsive to changing goals and context, and (iii) subject to consent, override, and adjustable autonomy. This reflects long-standing guidance that automation should act as a collaborative partner rather than an invisible controller^[22,87].

We define three role patterns in this overlay as *Choice Architect (R4)*, *Advisor (R5)*, and *Agentic Worker (R6)*, reflecting increasing initiative in shaping the decision environment, explaining trade-offs, and executing actions, but always under user oversight, reversibility, and the co-determination principles (T.A.N.) introduced earlier.

5.1 Role pattern R4: Subtle guidance in choice (AI as choice architect)

At R4, the AI begins to shape the *decision context* rather than the decision itself. As a *Choice Architect*, it uses small changes in salience and friction to make goal-consistent options easier to notice and compare while leaving selection entirely with the user. This draws on classic choice architecture and nudging, but under a stricter co-determination constraint: the system may guide attention, but must not covertly redirect goals or exploit vulnerabilities^[133–135]. In XR, this can be enacted through lightweight perceptual cueing, for example, gently highlighting items that match the user’s stated dietary goal in an AR aisle^[56], or rendering a

user-preferred route as more visually salient via in-view AR guidance while leaving all alternatives selectable^[136,137].

Mechanistically, this role pattern operates by re-weighting attentional evidence: the interface makes some cues more precise (more noticeable, easier to act on) so that acting on existing intentions requires less search and self-control. Because the same mechanism can become manipulation, R4 should be treated as scaffolding for autonomy, not behaviour steering. Self++ therefore binds Choice Architect nudges to co-determination principles (T.A.N.) safeguards: *Transparency* that the highlight is system-generated and why, *Adaptivity* that tracks the user's changing priorities rather than a single platform metric, and *Negotiability* through opt-out, adjustable strength, and consent for high-stakes nudges^[134,138]. These safeguards are especially important when R4 is running concurrently with Overlay 1 coaching, because the learner's heightened reliance and reduced situational bandwidth can otherwise make helpful layout indistinguishable from hidden coercion.

Finally, implementing Choice Architect support requires multi-objective reasoning: most real decisions trade off plural values (cost, safety, enjoyment, time), so the system should represent trade-offs and let the user steer weights rather than collapsing everything into an opaque score^[139]. In this way, R4 reduces decision friction and strategic uncertainty while preserving experienced authorship: the user can always recognise, contest, and revise how the system is shaping the field of choice.

This design stance also clarifies the ethical status of nudging within R4. The nudge debate has shown that whether a nudge is manipulative depends less on the inevitability of framing decisions and more on the mechanisms by which the nudging occurs and whether the direction of influence is transparent to the target^[48,134,135]. Self++ resolves this tension procedurally: every nudge in R4 must be transparently marked as system-generated and linked to the user's own stated goals, adaptively tuned to changing priorities rather than a fixed platform metric, and negotiable through opt-out, adjustable strength, and consent gates for high-stakes choices (Table 1). The system's influence therefore operates not by bypassing or subverting rational deliberation, but by scaffolding it, providing the informational and attentional conditions under which the user can deliberate more effectively while retaining full authorship over the resulting decision.

5.2 Role pattern R5: Informed deliberation (AI as advisor)

Where R4 shapes the choice environment, R5 externalises the deliberation *itself*. The AI becomes an Advisor: a conversational analyst that helps the user surface assumptions, compare futures, and reason through trade-offs, while keeping policy selection and endorsement with the user^[87,140]. This role pattern is especially important in contexts where persuasive optimisation can outperform genuine behaviour change in simulation but fail to translate into durable, owned decisions in the real world^[20,130]. In Self++, the Advisor is designed to feel like a co-determining voice that sharpens reflection rather than a persuader that steers outcomes.

Concretely, the Advisor provides *interactive evidence and counterfactuals* rather than a single “best” answer. It can assemble an XR dashboard that contrasts options across the user's stated criteria (for example, work-life balance, skill growth, risk, and social commitments), and allow the user to interrogate “why” and “what if” in place^[87,141]. A “day in the life” walkthrough, uncertainty bands, or side-by-side consequence traces can make long-horizon implications more legible without collapsing plural values into one score. The Advisor can also act as a memory and consistency check (“you previously prioritised family time”), and make second-order effects explicit (“skipping this meeting increases the chance of delaying Project X”) so the user is choosing with clearer foresight, not narrower freedom^[142,143].

R5, therefore, targets autonomy in its stronger sense: informed self-endorsement. It reduces “decision entropy” by illuminating unknowns and disagreements between objectives, but it must do so in line with the co-determination principles (T.A.N.). Transparency requires surfacing data provenance, assumptions, and uncertainty (and what the model cannot know). *Adaptivity* requires tuning explanation depth and modality to the user's expertise and momentary cognitive load. *Negotiability* requires editable goals, weights, and constraints, plus the ability to decline lines of reasoning, request alternatives, and override defaults. Together, these safeguards keep the Advisor supportive, legible, and revisable, so the user remains the author of the decision even when the AI is doing substantial analytic work^[47,87,140].

5.3 Role pattern R6: Empowered delegation (AI as agentic worker)

If R5 externalises deliberation, R6 externalises execution. Here, the AI becomes an Agentic Worker: it carries out well-scoped tasks on the user's behalf while remaining subordinate to the user's intent and oversight^[87,143]. The user delegates an outcome (and constraints), the AI proposes an executable plan, and the pair iterates until the plan is endorsed. This preserves autonomy because the AI's agency is not an independent authority, but an operational extension of the user's chosen policy.

Because delegation increases the risk of out-of-the-loop failures, complacency, and automation surprise, R6 requires explicit safeguards^[144,145]. Precisely, the Agentic Worker should operate as a proposal-approval loop: it presents what it intends to do (steps, assumptions, dependencies, and uncertainty), requests confirmation at appropriate checkpoints, and remains interruptible throughout^[87,146]. Intermediate autonomy is preferred over set-and-forget automation: maintaining user involvement at key junctures supports situation awareness and improves recovery when the environment deviates from expectations^[145,147].

Self++ implements these safeguards through the co-determination principles (T.A.N.). Transparency means the AI makes its intent, limits, and current authority legible (what it is doing, why, and what could go wrong). *Adaptivity* means autonomy is adjustable and can be tightened or loosened as the user's confidence, task criticality, and context change (for example, more confirmations for novel or high-stakes steps). *Negotiability* means delegation is always explicit, revocable, and renegotiable: the user can override, pause, or re-scope the task at any time, and the AI treats corrections as first-class inputs rather than friction^[87,148]. This keeps the system aligned with the user's values while reducing the need for persuasion; behaviour change is owned by the user because action follows endorsement, not covert steering^[20,130].

At the end of R6, Overlay 2 reaches its apex: the user experiences *augmented autonomy* in the strict sense; they remain the author of goals and approvals, while the AI reliably executes across tools and contexts with minimal cognitive burden^[87,143]. The result is higher throughput without surrendering control: autonomy is strengthened through delegation that is transparent, adjustable, and always negotiable^[144].

6. Overlay 3 – Societal and Existential Augmentation (Relatedness and Purpose)

Overlay 3 moves into the most aspirational domain of augmented agency: supporting relatedness, cultural embeddedness, and long-horizon coherence with values and purpose. If Overlay 1 supports “doing things right” (competence) and Overlay 2 supports “doing the right things” (autonomy), Overlay 3 supports “being the right self” in relation to others: sustaining relationships, repairing misunderstandings, and maintaining value-aligned trajectories in social and civic life. Importantly, these supports are not strictly sequential; in real-world settings, like project teams or classrooms, competence-building and autonomy often unfold within the context of social coordination and conflict. Overlay 3 is therefore a high-scope overlay that stabilises shared meaning and belonging while Overlays 1 and 2 operate in parallel.

Mechanistically, Overlay 3 targets uncertainty at the level of shared generative models. Teams and communities function best when participants converge on shared mental models, a mutual understanding of “what is going on” and “who is responsible for what”^[149–151]. At larger scales, collective sensemaking under information overload becomes a coordination problem where actors must distinguish signal from noise to avoid fragmentation^[152]. In SDT terms, this overlay protects relatedness by reducing social misattunement; in FEP terms, it minimises “societal” and “existential” entropy by ensuring that interacting agents remain aligned across both immediate actions and long-term timescales of meaning^[16,32].

Accordingly, Overlay 3 defines three role patterns mapping to R7–R9: *Contextual Interpreter* (making identity, norms, and downstream impacts legible to prevent social surprise); *Social Facilitator* (nurturing shared understanding and constructive conflict repair); *Purpose Amplifier* (supporting value-aligned self-regulation and life coherence to prevent value drift)^[139,143]. Because these role patterns touch the core of identity, co-determination is non-negotiable. The system must act as a user-legible partner, not a hidden governor. We therefore apply the co-determination principles (T.A.N.) as a hard constraint, requiring explicit transparency and negotiability whenever the system intervenes in relationships, values, or civic judgement^[87,130].

6.1 Role pattern R7: Big-picture contextualisation (AI as contextual interpreter)

R7 addresses a recurring failure mode of hybrid XR–AI settings: people can act locally (and fluently) while lacking context about identities, roles, norms, provenance, and downstream consequences. The *Contextual Interpreter* augments the user with situational and value-relevant legibility across two fronts: it surfaces information that may carry ethical, social, or practical significance for the user, without presupposing which normative framework applies. What counts as value-relevant is shaped by user configuration, cultural context, and the co-determination principles (T.A.N.), ensuring that context augmentation functions as epistemic support, expanding what the user can notice and anticipate, rather than as moral instruction.

On the social side, the Interpreter enforces *identity and role clarity* in mixed human–AI ecologies. In XR meetings or co-learning scenarios, it should make agent identity and function legible (for example, persistent labels or outlines that distinguish humans from AI agents and indicate the current role, such as “AI facilitator” or “human lead”). This is not cosmetic: disclosure cues help users calibrate expectations and preserve trust when agency shifts, including hand-offs structured via *HAT Swapping*^[19]. Evidence from AI service contexts suggests that identity disclosure can measurably shape user trust and uptake, reinforcing the need for explicit signalling rather than ambiguity^[153]. Beyond identity, the Interpreter can recover social signals that are weakened in mediated interaction (for example, shared gaze or attention cues), supporting mutual awareness and coordination^[97,154].

On the world side, the Interpreter bridges micro-actions to macro-consequences without coercion. It can surface value-relevant context that would otherwise be invisible at the moment of choice (for example, lifecycle or stakeholder impacts, long-horizon trade-offs, or alignment with the user's stated commitments), framed in the user's own value language where appropriate. This is particularly relevant when users confront large-scale, contested topics (public sentiment, politics, international relations), where sensemaking is shaped by information overload and social dynamics. Here, the Interpreter should act as a *plural-contextaid*: exposing uncertainty, provenance, and credible alternative interpretations rather than optimising for a single persuasive outcome^[130,152]. In Self++ terms, the Interpreter reduces “social surprise” and “contextual regret”: fewer breakdowns caused by misidentifying

counterparts, misreading norms, or discovering too late that an action conflicted with one's endorsed values.

Co-determination requirements are strongest in R7. *Transparency requires* identity disclosure, provenance cues, and uncertainty communication; *Adaptivity* requires tuning context density to attention and stakes (and backing off when low-value); and *Negotiability* requires user control over what contexts are surfaced, when, and at what sensitivity, including opt-out and override. Together, these safeguards ensure that context augmentation functions as user-aligned sensemaking support rather than covert social steering^[20,87].

6.2 Role pattern R8: Facilitating social connection (AI as social facilitator)

R8 moves from making context legible (R7) to actively improving how people relate and collaborate. Because real-world work, in classrooms, multidisciplinary teams, or cross-cultural communities, is inherently social, this role pattern often runs in parallel with competence-building (R1–R3) and autonomy support (R4–R6). Here, the AI acts not as a private companion, but as a light-touch facilitator that strengthens human-to-human coordination. By using XR to surface otherwise-missed social signals, it reduces the small misunderstandings that typically accumulate into conflict^[96,154].

The Social Facilitator builds shared mental models by restoring the attentional and intent signals often lost in mediated interaction. XR research demonstrates that cues such as gaze visualisation and mixed-reality communication markers can significantly improve grounding and social presence^[57,155,156]. The Facilitator extends this by visualising group dynamics, such as participation balance or conversational rhythm, allowing teams to self-correct without a human moderator^[79,157]. In fast-moving or jargon-heavy environments, the AI maintains common ground through optional micro-clarifications and role reminders, ensuring that shared understanding is actively supported rather than merely assumed^[24].

Where friction arises, the AI defaults to process support, summarising viewpoints and prompting perspective-taking, rather than adjudicating outcomes. This focus on conversational flow is critical, as fast responsiveness is tightly linked to felt connection^[158]. Crucially, Self++ treats R8 as explicitly pro-social: it aims to increase human-to-human contact rather than becoming the user's primary relationship. While AI companions can reduce loneliness by making users feel “heard”^[159], they also pose a risk of social drift toward synthetic companionship^[98]. The Social Facilitator mitigates this by preferentially scaffolding real-world relationships, inviting others in and encouraging repair after ruptures, and fading its own presence as human ties strengthen.

Because R8 touches group power and identity, it must be constrained by the co-determination principles (T.A.N.). *Transparency* requires clarity about what social signals are being sensed (e.g., “is the AI tracking my tone?”) and how feedback cues are generated. *Adaptivity* requires that the system can “read the room” and enter a do-nothing state when the group is thriving, and intervention would be intrusive. *Negotiability* must be socially contextualised, moving beyond individual consent to collective agreement. Specifically, *Collective Negotiability* should offer: *Mutual* opt-in, shared visualisations (like participation heatmaps) only appear if all members consent; *Privacy-by-role*, allowing individuals to opt out of certain group metrics without social penalty; and *Adjustable* mediation, enabling the group to negotiate the facilitation sensitivity, deciding, for example, whether the AI should flag interruptions or stay silent during heated creative debates. By situating negotiability within the group, the AI remains a tool for team co-determination, preventing failure modes like perceived surveillance or inadvertent shaming via data, and ensuring it remains a partner in human attunement.

6.3 Role pattern R9: Aligning life and values (AI as purpose amplifier)

R9 is the most delicate form of augmentation: the AI supports the user in living consistently with their self-endorsed values over long horizons, closing the gap between “the life I intend” and “the life I drift into”. This targets long-timescale misalignment (chronic regret, value drift, attention capture) that can accumulate into “existential surprise”. In SDT terms, the aim is not compliance but sustained autonomous self-regulation, where behaviour is owned and integrated rather than externally pressured^[32,160]. In FEP terms, the *Purpose Amplifier* helps the user maintain stable high-level priors (values and identity commitments) while flexibly updating mid-level plans and habits, reducing long-run expected free energy by steering away from trajectories the user later disavows^[16,44].

XR matters in R9 because immersive systems can intervene on the evidential stream that updates self and social priors and therefore can reshape what the user comes to expect of themselves and others. Self-representation effects make this concrete: embodiment can shift attitudes and self-models in ways that generalise beyond the session (e.g., reductions in implicit bias following avatar embodiment)^[161]. R9 interventions also often rely on *nudges-in-narrative* (for example, story-consistent exit cues or value-aligned prompts embedded in a virtual routine). Here, coherence becomes an ethical boundary condition: if users cannot distinguish evidential cues from narrative framing, persuasion risks becoming covert control, even when intentions are benevolent^[69]; these risks sharpen in high-realism XR, motivating stronger safeguards for long-horizon behavioural shaping^[162]. At the same time, XR can make long-horizon consequences and value conflicts perceptible rather than abstract: whereas R8 strengthens relationships and group functioning in situ, R9 uses XR as a value-facing *perceptual regulator* that externalises future selves, counterfactuals, and downstream impacts so the user can more reliably predict trade-offs and enact self-endorsed commitments^[163]. For example, immersive encounters with age-progressed future selves can shift intertemporal choices towards long-term benefits^[164]; VR

perspective-taking can change social attitudes and prosocial tendencies by making another standpoint experientially salient^[165]; and immersive climate experiences can improve learning and, in some settings, influence behavioural intentions and engagement by rendering invisible dynamics (e.g., ocean acidification) into lived evidence^[166]. These are not prescriptions of “what to value”; they are epistemic interventions that expand what the user can notice, anticipate, and contest, so value-consistent self-regulation becomes easier to sustain.

Practically, the system makes value-relevant discrepancies legible and actionable without turning them into coercion. It can surface periodic, user-configured reflections (e.g., how time, relationships, learning, and health track relative to stated priorities) and offer consensual, adjustable interventions aligned with the SPINED spectrum described earlier^[73]. The default is the least forceful effective move: inform, nudge, or entice before deter, suppress, or punish. This matters because heavy-handed control risks undermining autonomy even when it improves short-term behaviour^[32]. The design centre of gravity remains co-determination: the AI functions as another voice in the user’s deliberative ecology, amplifying what the user has already endorsed, not substituting its own normative agenda^[87].

What changes in R9 is that co-determination expands beyond the individual human–AI dyad to the *socio-technical loop*^[167] that shapes the dyad. R8 focuses on strengthening relationships and group functioning in situ; R9 governs the longer-run co-evolution of self, AI, and society by managing how systems shape preferences, norms, and incentives over time. Because XR interfaces can couple identity, attention, and affect into persuasive world-building, long-horizon alignment must treat recommendation and narrative loops as a coupled control problem: individual values guide system behaviour, system behaviour reshapes individual and collective priors, and institutions set reward structures that guide systems. This role makes such loops visible and steerable across two levels: personal settings for individual agency, and shared governance for teams and communities.

In R9, the co-determination principles (T.A.N.) become a societal as well as personal constraint, and they must be sharper than in earlier roles because the intervention surface now includes identity cues, narrative framing, and institutional incentives:

- **Transparency (reasons, framing, and incentives):** Interventions must include explicit “because” links to user-endorsed values (and visibility into what data is used, what is inferred, what incentives are optimised, and what uncertainty remains)^[20,87,130]. In XR, transparency also requires *framing legibility*. Users should be able to inspect when a cue is narrative scaffolding versus evidential guidance, because coherence failures can become ethical failures^[69,162].
- **Adaptivity (internalisation, not outsourcing):** The system should learn which supports feel autonomy-supportive (vs controlling), tune intensity and timing, and deliberately *fade scaffolds* so the user internalises routines rather than outsourcing self-regulation indefinitely^[32]. In XR, adaptivity also means calibrating how strongly self-representation or narrative devices are used, since these can update priors about self and others^[161].
- **Negotiability (contestable boundaries and escalation control):** Override must extend from moment-to-moment control (“not now”, “ask first”, “reduce frequency”) to *contestability*: users can challenge inferences, disable classes of interventions (e.g., identity-shaping cues, affective nudges), and demand alternative framings or evidence, with pathways for collective contestation when systems affect communities^[168]. Any escalation in intrusiveness (towards deter/suppress) requires explicit opt-in and reversible settings^[73].

These safeguards are also stability conditions against metric pathologies. When proxies become targets, they invite distortion and strategic behaviour (by systems and by users), captured by Goodhart’s and Campbell’s laws^[169,170]. R9, therefore, avoids single-score optimisation (e.g., “screen time” alone) as the governing objective. Instead, it treats wellbeing as plural, revisable, and context-dependent, keeping the user (and, when appropriate, the group) in the loop of redefining what counts as success^[87,139]. This closes the ethical loop: co-determination is not only a usability preference but a long-horizon alignment requirement.

Finally, R9 ties Self++ back to Dependent Origination: the self is not fixed but co-arises with conditions, including tools, social relations, and institutions^[171]. XR and AI become part of the causal web that shapes habits, identities, and norms; in turn, users and communities shape the objectives, feedback signals, and reward structures that shape AI systems. Read this way, Self++ is a co-evolutionary claim: *self evolves, AI evolves, society evolves*. Without transparent reasons, adaptive sensitivity to context, and negotiable boundaries, the coupled system is prone to drift into manipulation, capture, or adversarial gaming. With the co-determination principles (T.A.N.), R9 aims at a constructive blurring of boundaries: not loss of agency, but authentic agency amplified, where the user becomes more competent, more autonomous, and more connected over time, and where the socio-technical ecology is steered towards those human ends rather than away from them.

7. Self++ Design Propositions and Evaluation Checks

7.1 The propositions

Self++ is presented as a conceptual framework designed to be actionable, not only a taxonomy, but a set of commitments precise enough to be tested, debated, and refined through empirical work. In HCI and design-oriented research, frameworks become more

reusable when they are articulated as explicit claims that others can inspect, debate, and evaluate across contexts, rather than only described narratively. This aligns with interaction-design arguments for making knowledge transferable through concrete representations and critique, accounts of intermediate-level knowledge that support reuse and cumulative learning across projects^[172–174], and perspectives that emphasise articulating implications and principles so designs can be examined beyond a single instantiation^[175–177]. It also aligns with Human-Centered AI arguments that safety, control, and responsibility must be operationalised as design requirements rather than stated only as values^[45].

Accordingly, we state a compact set of propositions that summarise what Self++ claims about human–AI coupling under SDT and FEP, and how to evaluate systems that aim to instantiate these claims. We present these propositions as falsifiable design hypotheses, not validated findings. Each states a necessary condition that Self++ predicts must hold for co-determined augmentation to succeed; Table 2 pairs each proposition with concrete evaluation checks so that future empirical work can confirm, refine, or reject them. This distinguishes the propositions from heuristics or best-practice recommendations: they make specific, testable commitments about how SDT needs, FEP dynamics, and T.A.N. constraints interact under concurrent overlay activation, and they are intended to be wrong in instructive ways if the underlying theory is incomplete.

7.2 Empirical anchoring of the propositions

The propositions in Table 2 are offered as falsifiable design hypotheses. Several draw indirect empirical support from existing XR and HAT research, including work conducted in the author’s lab. This section maps each proposition to its current evidential status, distinguishing direct support (evidence from studies that test the specific claimed relationship), indirect support (evidence from analogous contexts that corroborate the underlying mechanism), and open hypotheses (claims that remain untested but generate concrete experimental predictions). This mapping is intended to guide future evaluation priorities and to make the framework’s empirical commitments transparent.

P1, Concurrency: *overlays act concurrently and can interfere. Indirect support.* XR-LIVE^[17] showed that learners in asynchronous shared-space virtual laboratory demonstrations used spatial-temporal assistive toolsets under conditions involving attention management, co-presence, and task guidance, highlighting trade-offs around cognitive load and split attention. The “Virtual Triplets” framework^[18] similarly introduced a mixed synchronous/asynchronous VR collaboration setting in which physical task execution and agent-mediated instructional coordination co-occurred, making concurrent overlay demands salient. Dong *et al.*^[56] further showed that AI-driven visualisation techniques in XR shaped decision-making under different levels of user autonomy, suggesting that concurrent perceptual and deliberative supports may interact. *What remains untested:* systematic manipulation of overlay combinations to measure specific interference patterns and recovery times.

P2–Timescale Alignment: *SDT needs map to uncertainty targets across temporal horizons. Indirect support.* Yang *et al.*^[74] found that AR-assisted construction assembly with low-agency control reduced workload while also reducing perceived autonomy, showing that gains in immediate task support can come at the cost of autonomy-related experience across different design horizons. Yousefi *et al.*^[24] measured human–AI team dynamics across confidence, satisfaction, accountability, and task performance, supporting P2’s claim that evaluation should extend beyond short-term task success to include interaction-quality outcomes. *What remains untested:* longitudinal studies tracking competence, autonomy, and relatedness indicators across their respective short-, intermediate-, and long-horizon commitments within a single deployment.

P3, Inspectability: *legitimate augmentation requires an inspectable, contestable AI voice. Indirect support.* The XAIR framework^[80] provides design evidence that AR explanations can support user agency when explanations remain accessible and are typically user-triggered. HAT Swapping^[19] further suggests that explicit disclosure of role and agency changes during hand-offs helps users calibrate reliance when control shifts between human and agent. Jing *et al.*^[97] demonstrated that bi-directional gaze visualisation in collaborative AR made attentional states more legible and improved coordination, providing an Overlay 3 analogue in which greater inspectability supports shared understanding. *What remains untested:* controlled experiments testing whether users can accurately identify what was influenced, why, and how to undo it, as specified in Table 2.

P4, T.A.N. Scaling: *co-determination strength must scale with scope and initiative. Currently a testable hypothesis.* No existing study systematically varies T.A.N. strength across overlays. However, work on trust calibration^[127,128] shows that the consequences of miscalibrated trust become more serious as systems take on more autonomous and consequential roles, which is consistent with P4’s prediction that higher-scope roles require stronger transparency and negotiability. Yousefi *et al.*^[89] further show that embodied virtual agents can elicit prosocial responses and that these effects depend on social-cue design, suggesting that socially and relationally scoped interventions may require stronger safeguards when they influence user behaviour. *Evaluation priority:* comparative studies testing whether T.A.N. safeguard strength scales appropriately from Overlay 1 through Overlay 3 as scope and initiative increase.

P5, Transition Legibility: *shifts in agency between role patterns must be perceptible and reversible. Direct support.* HAT Swapping^[19] is the most directly relevant study: it investigated how virtual agents act as stand-ins for absent human instructors in virtual training, showing that continuity cues and explicit disclosure of identity and role changes are important when agency shifts between human

and agent. Zhang *et al.*'s Virtual Triplets [18] further highlighted how mixed synchronous-asynchronous collaboration can introduce ambiguity about current agency and coordination. Han *et al.*^[79] similarly explored mediation by embodied virtual agents in triadic collaborative decision-making, showing that agent-mediated interaction can affect group coordination quality. *What remains untested:* systematic comparison of implicit (ambient) versus explicit (announced) transition cues and their effects on situation awareness and reclaim-time, as specified in Table 2.

P6, Endorsement over Compliance: *autonomy support preserves authorship over revision, not mere compliance. Indirect support.* Doudkin *et al.*^[20] provided critical negative evidence that persuasion effects predicted in synthetic and simulated participants did not translate cleanly to human pro-environmental behaviour change, highlighting the gap between surface persuasive success and genuinely internalised human uptake. This motivates P6's requirement that autonomy support must produce self-endorsed outcomes rather than surface-level agreement. Yang *et al.*^[74] similarly found that reducing user agency in AR assembly reduced cognitive workload but also reduced perceived autonomy, suggesting that support that makes action easier can still undermine the ownership that SDT identifies as necessary for sustained motivation. *What remains untested:* direct comparison of nudged versus unnudged conditions, measuring endorsement quality (e.g., whether users can explain "because..." in terms of their own goals and values) alongside performance.

P7, Collective Negotiability: *relatedness support requires shared-model alignment and group negotiability. Indirect support.* Han *et al.*^[79] showed that agent-mediated triadic collaboration shaped group dynamics and perceived collaboration quality, indicating that AI interventions in social settings operate at the group level and therefore require more than individual consent alone. Piumsomboon *et al.*^[155] found that sharing awareness cues in collaborative mixed reality improved grounding, performance, and usability, highlighting both the value and the design sensitivity of making social signals visible in multi-user settings. Together, these findings support P7's claim that relatedness support must be negotiated collectively, including decisions about what signals are shared, with whom, and under what conditions. *What remains untested:* studies testing whether participants can contest aggregation rules and thresholds while still collaborating smoothly, and whether opt-out mechanisms function without social penalty.

P8, Governance Contestability: *long-horizon alignment is socio-technical and requires contestation pathways. Currently a testable hypothesis.* No existing XR study has tested governance contestability as defined here. However, the broader literature on dark patterns^[131,132] and ethical nudging^[138] shows that systems can steer users toward unintended decisions when influence is insufficiently transparent, contestable, or autonomy-preserving, indirectly supporting P8's necessity claim. Piumsomboon *et al.*^[73] proposed the SPINED spectrum for XR disengagement based on expert elicitation and a preliminary online survey, providing a concrete example of how escalation pathways can be conceptually structured and comparatively assessed. This is directly relevant to R9's claim that increases in intervention intrusiveness should be governable, reviewable, and, in high-stakes contexts, explicitly consented to. *Evaluation priority:* longitudinal field studies testing whether users can effectively challenge the value assumptions and optimisation targets behind system recommendations.

The nine role patterns synthesise established concepts with novel contributions. R1 (Tutor) and R2 (Skill Builder) draw directly from well-validated instructional design and motor-learning literatures^[34,99,103,106-109]. R3 (Coach) extends these with XR-specific mechanisms (e.g., fault injection, overlay removal, and disclosed hand-offs) that have partial empirical support^[19,121,122]. R4 (Choice Architect) applies established nudging theory^[133-135] under novel co-determination constraints. R5 (Advisor) and R6 (Agentic Worker) are grounded in mixed-initiative, adjustable-autonomy, and human-centred AI literatures^[87,93,94], but their specific Self++ formulations (e.g., proposal-approval loops with T.A.N. constraints) are novel. R7–R9 are the most exploratory: they extend established ideas such as shared mental models^[149], social mediation, and long-horizon value support into XR-AI contexts where direct empirical validation remains limited.

8. Exemplary Scenarios of Self++

Meet Alex and Brooke, two 20-year-old university students facing the same three parallel demands: excelling in education, managing a part-time job, and maintaining a social life. Alex is steady and planful; Brooke is bursty and inconsistent. Brooke often stays up late gaming, wakes late, and misses classes, yet can become exceptionally creative and effective under pressure when they enter a flow state. Both adopt the Self++ XR system: an intelligent virtual assistant that runs mainly through XR glasses running in AR mode in everyday routines and switches to VR for immersive practice, stress relief, or structured reflection.

Crucially, Self++ is not a linear ladder. Its three concurrently activatable overlays can combine Role Pattern R1-R9 under co-determination (T.A.N.) (Transparency, Adaptivity, Negotiability), scaffolding growth while reducing uncertainty without taking ownership away from the user. This scenario uses a parallel timeline to show how the same nine role patterns can be adapted differently: for Alex, sustaining accumulation and calibration under pressure; for Brooke, enabling re-entry, preventing avoidable chaos, and preserving creative identity while building minimal, sustainable structure.

8.1 Building competence in education (tutor mode)

Morning, 8:30 AM (Alex) / 11:30 AM (Brooke): Alex heads to a chemistry lab for a new topic. Self++ enters Tutor (R1) and creates a safe,

learnable corridor: directional cues, relevant equipment highlights, and step-gated safety procedures. As Alex measures chemicals, the system offers immediate, gentle corrections. The effect is twofold: early attributable success (SDT competence) and lower surprise (FEP), so anxiety drops.

Brooke wakes late, already behind, and is at risk of avoiding altogether. Self++ still uses *R1*, but with a different aim: *re-entry*. Instead of a full lesson, it compresses the task into the smallest viable corridor (“two actions only”) and reduces shame-driven uncertainty by making the next step unambiguous. If Brooke attends the lab after missing prior sessions, Tutor mode prioritises *error prevention* and *safety gating* (what must not be missed) while keeping the interaction non-moralising and easily skippable. The goal is not discipline; it is enabling engagement in the first place.

Afternoon, 2:00 PM (Alex) / 3:30 PM (Brooke): A calculus assignment is due. Self++ shifts into **Skill Builder (R2)** and launches a VR practice module with an interactive whiteboard and immersive 3D visualisation. For Alex, it adapts difficulty on the fly and provides hints only after allowing time to think, keeping effort owned rather than outsourced. When Alex stalls, it uses subtle cueing (e.g., lightly highlighting a relevant formula) as a memory prompt rather than a solution dump. Support fades as proficiency stabilises. Alex experiences repeated, attributable successes, with challenges calibrated to avoid boredom or collapse into frustration.

For Brooke, R2 is structured as short, *variable sprints* rather than long drills. The VR module reframes practice as micro-gamified challenges that preserve novelty while still training fundamentals. Hints remain optional and late, and the system schedules practice windows where Brooke is most likely to reach flow. The system builds competence by capitalising on Brooke’s burst capacity, while quietly improving generalisation by varying contexts and constraints across sprints.

Evening, 7:00 PM (Alex) / 11:30 PM (Brooke): Approaching a mid-term test, Self++ becomes a **Coach (R3)**. For Alex, it overlays a heatmap on solutions (strong reasoning vs weak steps) and introduces metacognitive prompts. When it detects rushing through familiar sections, it nudges assumption checks. When Alex spirals after seeing peers post “10-hour study days”, Self++ shows a private progress dashboard that grounds self-assessment in Alex’s own trajectory rather than distorted social comparison. Coaching here trains resilience and calibration, preventing expertise from drifting into complacency or discouragement.

For Brooke, R3 targets a different brittleness: competence that appears mainly under adrenaline. Coach mode runs *safe pressure practice* (timed scenarios, interruptions, missing information) and gives short debriefs focused on stabilising performance without extinguishing creative leaps. It adds a single guardrail against impulsive “clever” shortcuts (constraint checks) while protecting Brooke’s ability to improvise. The point is robustness: creativity that remains reliable when conditions change, not just when panic peaks.

8.2 Empowering autonomy at work (advisor mode)

Weekday, 9:00 AM (Alex)/12:00 PM (Brooke): Alex works part-time at a tech start-up. Self++ adopts **Choice Architect (R4)**: it shapes the decision context while preserving authorship. When Alex views the task board through AR, one or two tasks are gently highlighted because they match Alex’s growth goals and the team’s priorities. Alex can choose anything, but indecision costs less. Suggestions are transparently tagged as AI prompts, preventing the “helpful layout” from becoming invisible steering.

Brooke also benefits from R4, but the main risk is derailment by *micro-choices*. Self++ makes “tiny start” actions the easiest to select (one visible tile that launches a 5-minute setup), and only adds friction where Brooke has explicitly opted in (e.g., a second confirmation before late-night gaming on weekdays). This preserves autonomy while reducing avoidable uncertainty created by impulsive context switches.

Midday, 1:00 PM (Alex)/2:30 PM (Brooke): During a mixed-reality meeting, the team hits a snag. Self++ moves into **Advisor (R5)**. For Alex, it offers multiple options with brief justifications and visible uncertainties, rather than a single “best” answer. Alex contributes these as discussable alternatives, combining AI evidence with human judgement (team preferences, creative insight, organisational constraints). Trade-offs become legible rather than intimidating.

For Brooke, R5 is tuned for avoidance collapse. The system uses concise counterfactuals and concrete next steps rather than long explanations. It may show two short futures (“if you delay” vs “if you do 20 minutes now”) and reframe tasks in Brooke’s own value language (e.g., protecting creative identity by linking required work to personal projects). The system reduces decision entropy without converting into compliance pressure.

Evening, 5:00 PM (Alex)/6:30 PM (Brooke): As Alex becomes more capable, Self++ supports **Agentic Worker (R6)**: delegated execution under a proposal-approval loop. Alex sets boundaries in an AR dashboard: draft reports automatically, but require review before sending; triage emails, but never touch messages marked sensitive. The system executes routine work quietly, pings for approvals at defined checkpoints, and stays interruptible. Autonomy strengthens because delegation is explicit, scoped, and revocable, and Alex learns meta-autonomy: when to hand off and when to stay hands-on.

For Brooke, R6 is “anti-chaos delegation”: preventing administrative failure (missed emails, missed forms, missed replies) from consuming capacity and causing downstream social or institutional penalties. The system drafts messages and proposes schedules,

but preserves consent checkpoints for anything consequential. Delegation here protects autonomy by preventing small failures from snowballing into externally imposed constraints.

8.3 Fostering relatedness in social life (networker mode)

Friday, 7:00 PM (Alex)/8:30 PM (Brooke): Alex and Brooke are friends, and they are both meeting the wider group at a café. Alex is keen but socially anxious, especially with new acquaintances, while Brooke arrives later and is noticeably quieter than usual. Self++ foregrounds Overlay 3 support for both of them. At **Contextual Interpreter (R7)**, the system reframes the evening as legitimate recovery rather than “lost productivity”, showing Alex’s completed commitments and a simple view of the week’s balance. This reduces guilt and supports value-consistent wellbeing.

For Brooke, the challenge is often not anxiety but inconsistency: disappearing, then avoiding people due to embarrassment. R7 therefore makes consequences legible privately and without shame (e.g., “you have not replied to X; a short repair message prevents drift”). It also clarifies social and institutional context (“this message expects a reply today” vs “FYI only”), reducing social surprise.

At the café, Self++ provides optional, privacy-respecting cues: names and agreed-to “common ground” hints for introductions. It stays light-touch: enough to reduce awkward uncertainty without making either user dependent. As the conversation unfolds, it shifts to **Social Facilitator (R8)**. When Alex notices Brooke is quiet, the system supports *human-led inclusion* rather than stepping in as the social actor. In Alex’s view, it offers a gentle, non-intrusive prompt such as “Brooke has not spoken for a while; consider a check-in or an easy entry point” and surfaces a low-stakes bridge topic grounded in shared context (e.g., “ask about the design sprint they enjoyed”), without exposing private data. Alex uses this to invite Brooke in: a simple question, a shared joke, or an explicit acknowledgement (“glad you made it”) that lowers pressure.

For Brooke, R8 supports *repair and re-entry* without public call-out. In their view, the system can offer opt-in micro-supports: a one-tap “join-in” suggestion (two or three possible responses), a private reminder that silence is acceptable, and a quick “common ground” cue that helps them re-engage on their own terms. If miscommunication appears, Self++ can offer a neutral micro-summary of viewpoints to prevent escalation. If the group is thriving, it does nothing. Relatedness is strengthened by improving human-human coordination and continuity, not by replacing relationships.

Saturday, 10:00 AM (Alex)/1:00 PM (Brooke): The next day, Self++ runs a short **Purpose Amplifier (R9)** reflection. For Alex, in a calm AR ambience, it visualises how study, work, and social care link to longer-term aspirations. It reframes these as mutually reinforcing rather than competing: competence as foundation, autonomy as agency, relatedness as meaning and resilience. It may suggest value-aligned opportunities, but these remain invitations, not prescriptions. The purpose is coherence: reducing long-horizon drift by making values actionable without becoming coercive.

For Brooke, R9 protects creative identity while reducing drift that later feels like betrayal. The system does not prescribe “be disciplined”; it supports value coherence through opt-in, inspectable simulations of downstream consequences (e.g., a “future self” contrast between chaotic nights and minimal structure that preserves creative time). It keeps framing legible and editable, ensuring Brooke can rewrite narratives in their own language.

8.4 Balancing conflicts via co-determination (where Self++ earns its keep)

Life becomes convoluted when domains collide. During crunch week, Alex faces an exam, a critical work presentation, and a close friend’s wedding within two days. Stress spikes because each demand threatens another.

Anticipation and planning (weeks earlier): Self++ notices the clash early and nudges forward preparation: earlier study blocks, a VR practice exam, and protected time around the wedding. This is active uncertainty regulation: fewer last-minute surprises mean less stress.

Negotiating autonomy (when work shifts): Alex’s boss asks to move the presentation to the wedding day. Self++ (*R5-Advisor*) generates a private XR comparison of two timelines and their consequences. It finds feasible alternatives (another slot, coverage options) and helps draft a professional email proposing a solution. Alex remains the author; the system makes negotiation easier and less threatening. The boss agrees to reschedule.

Dynamic rebalancing (day-of): The exam and wedding still share a day. Self++ shifts roles fluidly: R2–Skill Builder/R3–Coach at dawn (focused VR review on weak areas), R4–Choice Architect/R5–Advisor before the event (logistics checks and timing nudges), R8–Social Facilitator at the wedding (mostly silent, with optional translation subtitles for an overseas relative), and recovery that night (a short VR calming session).

Brooke’s conflicts often look different but are equally entangled: late-night flow collides with a Monday deadline, while a friend asks for help moving flat on Sunday morning. The system does not “optimise” Brooke; it makes the conflict legible and recoverable. A minimal co-determination response may combine: (i) *R4* friction only where Brooke opted in (confirming the cost of starting another game), (ii) *R5* two short futures (help friend + miss quiz vs delay help by 90 minutes and keep both), (iii) *R8* a repair message draft (“I

can help at 9:30; compulsory quiz at 8”), and (iv) *R6* alarms and a checklist, with approvals at key points. Brooke still chooses; the system reduces avoidable surprise and supports agency-preserving recovery.

In both cases, success is not that Self++ “won” the trade-off, but that it helped keep all three SDT needs in view under pressure, while making interventions transparent, adaptive, and negotiable.

8.5 Outcome: A co-determined growth trajectory

Across domains, Self++ scaffolds without taking the steering wheel. In education, it builds competence from onboarding to robust mastery while training calibration: for Alex, steady accumulation and bias-resilient self-assessment; for Brooke, re-entry corridors, sprint practice, and pressure-safe robustness that protects creativity. At work, it strengthens autonomy from gentle prioritisation to explicit, reversible delegation: for Alex, throughput with oversight; for Brooke, anti-chaos delegation that prevents small failures from becoming externally imposed constraints. In social life, it reduces social uncertainty, supports repair, and deepens coherence with values and purpose: for Alex, confidence and presence; for Brooke, continuity and reconnection without shame.

When conflicts arise, the Overlays overlap rather than queue: competence support can run during autonomy negotiation inside a social obligation. Throughout, *T.A. N* keeps augmentation legitimate: users can tell what is guided and why, support adapts and fades with growth, and overrides or renegotiations remain always available.

The result is not an overnight transformation but a sustainable trajectory: both users become more capable, more self-directed, and more connected, recovering quickly from mistakes because the system catches “just enough” to get back on track and then steps back.

9. Discussion

Self++ responds to recurring XR-AI issues by treating *co-determination* as a design requirement rather than a usability feature. Below, we consolidate the main implications into three themes: (i) agency and calibration, (ii) ethical boundary conditions for experience-shaping systems, and (iii) institutional and governance implications.

9.1 Agency, calibration, and metacognitive accuracy

A central risk in XR-AI assistance is **erosion of agency**: systems can take control “for the user’s benefit,” undermining learning, ownership, and accountability. Self++ counters this by keeping the human as the author of action while the AI scaffolds performance and decision-making. In *T.A.N.* terms, *Negotiability* operationalises consent, override, and renegotiation so assistance remains revocable and role boundaries stay explicit. This aligns with coactive teamwork, where human and AI remain interdependent partners rather than a controller and a controlled system^[148], and with mixed-initiative design that treats initiative shifts as coordination problems rather than hand-offs to be hidden^[93]. A practical expectation is fewer mode-confusion episodes and fewer “why did it do that?” moments because intent and authority are made legible before the system acts.

A second challenge is **calibration**: users must calibrate both *trust in the system* and *confidence in themselves*. Poorly designed systems invite over-trust (misuse) or under-trust (disuse), undermining human-AI teaming^[178]. XR can amplify these errors: immersive guidance can inflate perceived competence, while a single failure can collapse trust. Self++ addresses this through *Transparency* and *Adaptivity*: the system should disclose capability limits, intent, and uncertainty, and adjust autonomy as the user and context change. Clear uncertainty and rationale cues support appropriate verification^[128] and can improve satisfaction, situation awareness, and team performance^[24] by narrowing the “gulf of evaluation” between user expectations and system behaviour^[77].

Self++ also treats **self-assessment biases** as part of calibration. Novices can overestimate mastery while experts underestimate gaps; XR training that maximises ease can worsen these illusions by confounding performance with assistance. Self++ therefore emphasises calibrated feedback, scaffolded reflection, and guidance fading: the system should make the source of success legible (user skill vs AI help) and progressively withdraw support as competence stabilises. This aligns with evidence on prompting explanation and fading hints^[103], and with “desirable difficulty” accounts showing that structured challenge improves retention and reveals limits^[121,122]. More broadly, immersive representations can be used for reflective sensemaking when they externalise uncertainty structures and alternatives rather than presenting a single persuasive conclusion^[179]. When paired with plural perspectives in deliberation, these mechanisms can reduce cognitive illusions amplified by digital mediation^[5,6].

These agency and calibration risks compound when considered across the user’s full range of activities, because the same user will typically operate at different role-pattern levels across skill domains simultaneously. A professional might function at *R6* (Agentic Worker) for routine administration while operating at *R1* (Tutor) for a newly acquired technical skill and at *R3* (Coach) for a long-practised competence. This is not inconsistent; it reflects the established finding that competence is domain-specific^[99] and that appropriate support must track the user’s capacity and goals in each domain independently. Agency preservation, therefore, requires not only moment-to-moment legibility within a single role pattern but a portfolio-level view: the user should be able to inspect and negotiate which domains receive which level of support, ensuring that delegation decisions (*R6*) are informed by their own assessment of where competence lies and where effort is better invested in development versus delegation. Overlay 2 mechanisms,

particularly Choice Architect (R4) and Advisor (R5), serve this governance function by making cross-domain allocation explicit and revisable. This also has implications for calibration: if the system presents a uniform confidence signal across domains where the user's actual competence varies widely, it risks inflating self-assessment in weak areas and suppressing growth in strong ones. The relationship between competence support and autonomy support is therefore not merely temporal (first build skill, then exercise choice) but structural: autonomy mechanisms govern how competence support is distributed across the user's activities, and miscalibration at this portfolio level is itself an agency failure that T.A.N. is designed to prevent.

9.2 Ethical boundary conditions for experience-shaping systems

Because XR systems can shape the evidential stream, Self++ is not a moral optimiser and should not be framed as "making people good." Its goal is to help users act more consistently with what they already endorse, while keeping influence inspectable and revisable under T.A.N. This stance requires an explicit acknowledgement: Self++ is ethically procedural, not substantive. It does not encode a preferred moral framework or assume universal agreement on what constitutes a good life, a responsible choice, or a well-functioning community. Such standards vary across cultures, belief systems, and individual histories, and any system that hard-codes a single ethical orientation risks becoming an instrument of cultural imposition rather than support. Self++ addresses this by locating its normative commitments at the level of process rather than content. T.A.N. specifies how the system must behave, making influence visible, tuning support to context, and preserving the user's right to contest and override, without specifying what the user should value or endorse.

However, procedural safeguards are only as strong as their implementation, and each can fail in characteristic ways. Table 3 presents selected, illustrative examples, rather than an exhaustive taxonomy, by mapping each role pattern to a primary failure mode, the mechanism by which well-intentioned support can drift into harm, the T.A.N. safeguard intended to prevent it, and the residual risk that the safeguard itself may prove insufficient. This residual risk matters because transparency, adaptivity, and negotiability can themselves be undermined by workload, habituation, miscalibration, or strategic misuse. Three cross-cutting dynamics deserve particular attention: **escalation drift**, where support gradually increases in intrusiveness without renewed consent (especially R4→R6 and R7→R9); **Goodhart dynamics**, where single-metric optimisation corrupts the intent of support; and **adversarial gaming**, where T.A.N. transparency is exploited by users or third parties. When the system detects that its own support may be undermining autonomy, for example, when endorsement scores decline even as performance improves, it should surface the discrepancy transparently, offer to reduce support intensity or change mode, log the event for audit, and default to the least intrusive level if the user is unresponsive. In this sense, T.A.N. is not a one-off interface property but an ongoing governance commitment requiring monitoring, contestability, and periodic review.

Table 3. Failure-mode analysis: selected examples of how T.A.N. safeguards can fail and how well-intentioned support can drift into harm.

Role	Failure Mode	Mechanism of Drift	T.A.N. Safeguard	Safeguard Failure Risk
R1	Dependency: user cannot perform without guidance	Guidance never fades; early success is confounded with AI assistance, inflating self-assessment	Adaptivity: fade schedule linked to demonstrated competence, not elapsed time	Fade triggers are miscalibrated; system defaults to "safe" (more support) under uncertainty
R2	Skill brittleness: user performs well only under augmented conditions	Practice variability is insufficient; ghost tracks and shadow cues become permanent reference	Adaptivity: introduce controlled variability; withhold hints progressively	User opts out of challenge increases via Negotiability, inadvertently freezing development
R3	Mode confusion: user cannot tell whether human or AI is in control	Poorly communicated agency transitions; silent role swaps	Transparency: explicit disclosure of role and agency changes (HAT Swapping protocol)	Disclosure is technically present but not perceptually salient during high-workload conditions
R4	Covert steering: "helpful layout" becomes invisible manipulation	Nudges are not marked as system-generated; user cannot distinguish curated from neutral views	Transparency: mark all nudges; provide "unnudged view"	Users habituate to nudge markers and stop noticing them; salience degrades over time
R5	Anchoring bias: user over-relies on AI framing of trade-offs	AI consistently presents options in the same order or with the same emphasis; user adopts AI framing as	Negotiability: editable goals and weights; ability to request alternative framings	User lacks the domain expertise to recognise when the AI's framing is skewed

		their own		
R6	Out-of-the-loop complacency: user rubber-stamps proposals without genuine review	Checkpoint frequency is too low; delegation scope creeps without explicit renegotiation	Negotiability: explicit delegation scope; adjustable checkpoint frequency; revocability	Approval fatigue: too many checkpoints lead to routine approval without scrutiny
R7	Information overload or filter bubble: context density is too high or too narrow	System surfaces too much context (attentional overwhelm) or too little (false certainty)	Adaptivity: tune context density to attention and stakes; back off when low-value	Attention estimation is inaccurate; system cannot reliably predict when context is helpful
R8	Surveillance perception: participants feel monitored rather than supported	Social signal sensing is too granular or insufficiently disclosed; facilitation feels like performance management	Transparency: disclose sensing granularity; Negotiability: collective opt-in, privacy-by-role	Individual opt-out creates social asymmetry (those who opt out are perceived as uncooperative)
R9	Value imposition: system's inferences about user values are wrong or culturally biased	Inferred values reflect designer defaults rather than user-endorsed commitments; narrative framing covertly steers identity	Negotiability: contestable inferences; ability to disable intervention classes; governance hooks	User lacks vocabulary or confidence to articulate disagreement with inferred values

However, this procedural stance is not value-free. The decision to surface certain consequences rather than others, to frame trade-offs in particular ways, or to define what constitutes “drift” from endorsed values all involve normative assumptions that may reflect the designers’ cultural position. Self++ therefore treats these assumptions as first-class design parameters: they must be documented, auditable, and revisable through the negotiability mechanisms described in [Section 6](#), and through the participatory and co-design approaches noted in [Section 9.5](#). Where systems affect communities rather than individuals alone, collective contestation pathways (R8’s collective negotiability and R9’s governance hooks) become the primary safeguard against monocultural default assumptions. In short, Self++ aspires to cultural humility: it holds that the conditions for flourishing are real and important, but that their specific expression must be determined with users and communities, not for them. This distinction becomes especially important in higher-scope interventions (identity, relationships, civic judgement, long-horizon behaviour), where even well-intentioned support can become covert preference shaping.

Self++ also has a clear **dual-use** risk. The same mechanisms that scaffold autonomy and relatedness can be repurposed as manipulation, including persuasive dark patterns^[131,132] or “sludge” that preserves the appearance of choice while steering outcomes^[135]. Personalisation data can further enable profiling and undue influence. At minimum, safeguards should follow ethical nudging guidance^[138] and human-centred AI principles^[45]: transparency should include intent (who benefits), users should have meaningful opt-out and data control, and systems should be auditable for manipulative behaviour.

XR introduces an additional manipulation surface via **self-presentation**. Systems that filter or reframe a user’s social signals (e.g., making them appear happier or suppressing negative affect) can function as social dark patterns if users cannot inspect or contest the transformation^[180]. Even when users consent initially, default-on transforms risk identity drift and misattribution in consequential settings (work evaluation, conflict repair, health, legal contexts). A Self++-consistent constraint is: self-presentation interventions require high-salience disclosure, editable parameters, and easy reversion, with stronger safeguards as stakes rise.

A related policy gap concerns **state-aware assistance** when the user is plausibly impaired (drowsy, medicated, intoxicated, acutely stressed). State sensing can increase safety, but it also increases surveillance and paternalism risk. A Self++ pattern is to treat impairment detection as risk gating, not permission to seize control: disclose what is sensed and its reliability, shift to safer defaults (more confirmations, reduced autonomy, fewer irreversible actions), and require explicit, revocable opt-in for any escalation beyond nudges. In group settings, inferring impairment is sensitive, so sharing it outward should be prohibited by default except for clearly defined, consented safety protocols.

Finally, the next generation of XR will increasingly generate experience (adaptive soundscapes, affective ambience, personalised visuals, fully generative one-off environments). These can support restoration, creativity, and engagement, but also introduce covert mood steering and narrative capture. Self++ therefore treats **framing legibility** as a hard requirement: users must be able to distinguish evidence from aesthetic framing and persuasive scaffolding, and adjust or disable these overlays. This also applies to wellbeing applications such as mindfulness and self-transcendent experiences, which are promising but high-leverage; reviews suggest XR can support contemplative practice when interventions are bounded and autonomy-supportive^[181]. Here, the T.A.N

gradient matters: intent disclosure, adjustable intensity/frequency, and debrief mechanisms help users integrate benefits without dependency.

On the theme of **transcendence**, this ambition resonates with a Buddhist view that liberation becomes possible through insight into how experience is conditioned, classically articulated through dependent origination^[174]. In this account, ignorance is not a lack of knowledge but a structural error in perception: treating impermanent, interdependent processes as fixed and self-contained, including the construct of a stable, separate self^[35]. Contemplative practice aims to correct this error by making the causal chains between perception, craving, and habitual reaction visible and interruptible. Self++ shares this structural logic without claiming equivalence: by making the conditions of mediated experience transparent, adaptive, and negotiable, the framework supports the user's capacity to notice what is being shaped, by whom, and toward what ends. In this reading, co-determined XR-AI systems could function as attentional scaffolds that sustain the reflective clarity that contemplative traditions regard as a prerequisite to wise action, provided such systems remain bounded, autonomy-supportive, and subject to the user's ongoing consent^[181].

9.3 Self++ and diverse trajectories: Disability, delegation, and the reorganisation of self

Self++ is designed around progressive scaffolding, with support that fades as competence grows; however, this framing requires careful qualification because not all users are on the same trajectory, and not all trajectories point toward reduced AI involvement, as illustrated by three cases.

From a **disability and permanent support** perspective, assistive technologies are often experienced not as crutches but as extensions of the self, part of how a person acts in the world [cf. the extended mind thesis, 3]; a wheelchair user does not experience their chair as a temporary scaffold awaiting removal, and similarly, a person with a cognitive impairment may rely on R1-level guidance permanently, where this reliance represents appropriate, identity-consistent support rather than a failure to progress; Self++ accommodates this by anchoring adaptivity in the user's *endorsed goals and current capacity* rather than in a fixed developmental endpoint, such that if remaining at R2 (Skill Builder) aligns with what the user can and wants to do, that is a valid steady state, and the T.A.N. requirement here is that the system does not assume the user should progress, does not impose guilt or friction for staying, and continues to adapt support to changing circumstances within the user's chosen level of engagement.

Considering **elective delegation and the two paths of mastery**, even for users without disabilities, the relationship between competence and AI support is more nuanced than a simple reduction over time, as at higher levels of mastery two valid paths emerge, where some users seek to *fully embody* a capability and gradually need the tool less as they internalise the skill, while others seek to *leverage* the tool more precisely to reach outcomes they could not achieve alone; for example, a professional translator may hand off routine translation entirely to AI while investing their freed capacity in nuanced literary work that demands deep human judgement, and in this reading competence at the expert level is not about doing things without tools, but about knowing when and how to deploy them to serve higher-order goals, which Self++ supports through Overlay 2's negotiability mechanisms, allowing the user to explicitly choose "help me do this" (R2/R3 scaffolding toward embodiment) or "do this for me" (R6 delegation toward leverage), and to shift between them as context and priorities change.

Over time, through a **reorganisation of the self around AI**, what emerges is not simply less dependence on AI, but a redistribution of engagement, where in domains central to a person's identity and values, what might be called the "core self," users tend to engage with AI more critically and precisely, refining and deepening their interaction over time or choosing to embody the capability entirely on their own, while in more peripheral domains they delegate more freely, investing the reclaimed time and attention into what matters most; this reorganisation is consistent with SDT's distinction between intrinsic motivation (deeply owned, identity-consistent pursuits) and extrinsic regulation (actions motivated by external contingencies)^[32,160], and Self++ can support this by making the pattern visible, for instance through R9 (Purpose Amplifier), which could help users track which domains they are delegating, which they are deepening, and whether this distribution aligns with their endorsed values, without prescribing what the "correct" distribution should be.

Taken together, the **implications for T.A.N.** are not a weakening but a contextualisation of its requirements, such that adaptivity does not always mean fading but instead means responsiveness to the user's changing *relationship with the capability*, where for some users adaptivity may involve intensifying support when conditions deteriorate, and for others it may involve shifting the kind of support (from scaffolding to delegation infrastructure) rather than reducing its amount, with negotiability becoming especially important as users must be able to define their own trajectory, including the decision to remain at a given level of support indefinitely, without the system treating this as a failure state.

9.4 Beyond human-level reasoning: Self++ as an interface for superhuman and self-improving AI

A further motivation for Self++ is the plausible trajectory toward **artificial super intelligence (ASI)**^[10], including systems that improve via self-play, self-generated curricula, recursive self-improvement, or scalable oversight beyond direct human feedback. AlphaGo highlighted how learned policies can produce strategies that surprise experts, and AlphaGo Zero strengthened the point by reaching high performance with minimal human priors beyond the rules^[182]. In broader optimisation and scientific settings, analogous agents

may propose solutions in high-dimensional spaces that are useful yet difficult for humans to justify or even interpret. This creates an *interface* problem as much as a capability problem: when reasoning outruns ordinary human intelligibility, the risk is not only power, but loss of the ability to *understand*, *contest*, and *appropriately rely on* proposals, a concern central in control and alignment work^[10,183,184]. XR can exacerbate misplaced deference because immersive presentation can add authority while hiding assumptions and failure modes; however, it can also be part of the remedy by using human multimodal perception for higher-bandwidth sensemaking (spatialised structure, audio, haptics, thermal cues) so users can more fully *grok* complex trade-offs in parallel. Recent work on superalignment frames this as a weak-to-strong governance problem: humans are the weak supervisors who must still control systems much smarter than themselves, so interfaces like Self++ become part of the oversight stack that keeps decisions inspectable and contestable rather than merely convenient^[185].

Reframed in Self++ terms, superhuman reasoning raises the required strength of T.A.N rather than weakening it. *Transparency* must shift from “explain the answer” to “make the *decision structure* legible”: expose constraints, trade-offs, counterfactuals, and uncertainty in forms people can interrogate, aligning with interpretability aims of human-meaningful representations^[186,187]. *Adaptivity* must tune that legibility to human limits and stakes (what to surface now, what to defer, when to escalate evidence), while considering epistemic humility about boundary conditions and distribution shift^[183]. *Negotiability* becomes the core safety valve under asymmetric intelligence: even if the system can discover options humans would not find, *adoption remains co-determined* via explicit veto points, staged commitments, and contestable assumptions, echoing the motivation for scalable supervision and preference-based oversight while recognising their limits^[184,188]. In this reading, Self++ treats XR as a sensemaking overlay between human values and superhuman optimisation: advanced intelligence can be usable without becoming unquestionable, because T.A.N keeps proposals inspectable, adjustable to context, and always contestable under human authority.

9.5 Limitations and future work

Self++ is a role-based interaction theory, so its main limitations are less about conceptual coverage and more about operationalisation: building systems that deliver co-determined support reliably, measuring SDT-relevant states in situ, and validating effects over time and across contexts.

Operational feasibility in real-world XR. Running multiple roles as concurrently activatable overlays requires real-time policy arbitration, conflict handling, and fast failure recovery, and making automation behave as a true team player remains demanding in practice^[22,75]. Many interactions also assume robust sensing and timely feedback; current hardware constraints can break legibility cues or mis-trigger interventions, and response delays can degrade trust, coordination, and perceived social presence^[158]. A practical agenda is to specify role-specific tolerances (latency, sensing fidelity), then design graceful degradation paths when those tolerances are not met.

Interaction design of role-pattern transitions. Self++ specifies what should change when the system shifts between role patterns, the functional intent, support level, and T.A.N. requirements, but deliberately leaves underspecified how that change is communicated to the user in XR. When the system transitions from Tutor (R1) to Skill Builder (R2), or when Coach (R3) and Advisor (R5) activate concurrently during a team training scenario, the perceptual and interaction design of that transition, whether it manifests as a gradual fading of visual cues, an explicit notification, an ambient shift in soundscape or colour temperature, or a change in agent embodiment or behaviour, remains an open design research question. This omission is intentional: the appropriate transition idiom is likely to be highly dependent on modality (AR vs VR), task criticality, attentional capacity, and user preference, making premature specification counterproductive. However, that transition design is not merely cosmetic. Poorly communicated role shifts risk the mode confusion and automation surprise that Self++ aims to prevent (Section 4.3), while overly salient transitions may disrupt flow or impose unnecessary cognitive load. We therefore invite empirical investigation into transition legibility, including comparative studies of implicit (ambient) versus explicit (announced) role-shift cues, user-configurable transition salience, and the perceptual markers that best support situation awareness during concurrent overlay activation. Table 2, proposition P5, provides initial evaluation criteria for this work.

Measurement, legibility, and the cost of co-determination. Self++ presumes systems can tune support to competence, autonomy, and relatedness dynamics, yet reliable real-time indicators for these constructs remain limited. Trust and reliance have workable behavioural signals (for example, hesitation and overrides)^[127], but analogous indicators for competence frustration or relatedness quality are underdeveloped. Future work should develop lightweight in situ measures (micro-self-reports and unobtrusive multimodal signals) that are accurate enough to drive adaptation without becoming intrusive or surveillance-like. In parallel, designers must avoid over-scaffolding: persistent support can create dependency and out-of-the-loop problems^[145], and can inflate self-assessment when assistance is confounded with skill^[189].

While guidance fading and structured challenge offer principled countermeasures^[106,121], the right fade schedule is task- and person-dependent, so scaffolding schedules should be treated as testable design parameters, including whether calibration cues reduce inappropriate deference^[128]. Legibility also becomes harder at higher overlays: explaining context interpretation or long-horizon coaching requires user-meaningful and auditable accounts, but much XAI work targets local model decisions rather

than extended cognitive interventions^[141]. XR further constrains explanation because it must fit attention limits without breaking immersion; early AR-focused frameworks are promising but immature^[80]. Finally, transparency and negotiability introduce overhead: over-disclosure and frequent confirmations can interrupt flow and increase workload^[105], echoing classic concerns that excessive mode signalling can harm situation awareness^[124]. Future systems should therefore prioritise selective, stake-triggered transparency, quiet background cues, and user-tunable thresholds.

To reduce the gap between theoretical constructs and engineering implementation, a practical next step is to translate the co-determination principles into concrete programmable logic. For Negotiability, systems could incorporate a “Trust Protocol” built around pre-assigned autonomy budgets or risk thresholds. This would allow roles such as the Agentic Worker to execute low-stakes actions without constant confirmation, while reserving interruptions for exceptions and boundary conditions. Likewise, the timing of Adaptivity could be formalised via “Drift Triggers”: statistical thresholds (for example, a plateau in performance variance) that indicate reduced entropy and automatically prompt a transition from support (Skill Builder) to challenge (Coach). Finally, validation will require evolving the evaluation checks (Table 2) from qualitative audits into quantitative proxy metrics, such as tracking reclaim rates (the frequency of user overrides) to infer autonomy calibration, or measuring turn-taking balance to quantify relatedness quality in multi-agent settings. To be clear, the propositions in Table 2 are offered as empirically testable claims, not as design guidelines whose validity is assumed; progress requires that future studies treat negative or partial results as informative refinements of the framework rather than implementation failures.

Generalisability, integration, and evaluation infrastructure. Relatedness and autonomy are expressed differently across cultures and contexts, so Self++ needs stronger guidance on how interaction styles and boundaries should vary under different self-construals and relational norms^[36,37]. Because Self++ touches identity-, relationship-, and purpose-adjacent support, participatory and co-design approaches are important, particularly with marginalised groups who may face distinct risks and expectations^[175]. Implementation choices will also be shaped by AI capabilities: large multimodal models could expand context understanding, dialogue, and planning^[82], but raise controllability and presentation challenges, including how to surface uncertainty, provenance, and constraints in user-legible form^[83]. One pragmatic direction is hybrid architectures where foundation models are sandboxed or used offline, while real-time interaction is governed by stricter policies aligned with T.A.N. Finally, progress will be slow without shared benchmarks and compliance metrics: beyond outcome measures, we need longitudinal and teaming-quality evaluations, including when combined human–AI systems outperform human-only or AI-only baselines^[23], operational measures of team fluency^[149], domain norms for organisational settings^[167], and explicit metrics for T.A.N. compliance suitable for audit and comparison. These efforts align with human-centred AI guidance^[45] and emerging trustworthy agent frameworks^[146], and could provide a foundation for reproducibility and eventual standardisation.

Implementation requirements by overlay. Self++ assumes real-time sensing and response capabilities that vary in stringency across overlays, though all three benefit from advances in multimodal foundation models^[82]. Overlay 1 role patterns (R1–R3) are the most latency-sensitive, requiring tight perception–action loops for perceptual cue updates and feedback delivery, robust spatial tracking, and reliable object and action recognition. Current AR headsets approach these requirements for constrained task domains but remain limited in field-of-view, occlusion handling, and outdoor robustness. Even at this sensorimotor level, vision–language models can improve cue relevance, error interpretation, and context-aware feedback timing, complementing the spatial tracking layer with semantic understanding^[80,102]. Overlay 2 role patterns (R4–R6) operate at deliberative timescales and are therefore less latency-sensitive, but require accurate context modelling, reliable natural language interaction, and secure delegation infrastructure; controllability and uncertainty communication remain challenging^[83]. Overlay 3 role patterns (R7–R9) require longitudinal user modelling, cross-session memory, cultural and normative sensitivity, and reliable affect and social signal recognition, capabilities that remain immature in current systems, though fast and contextually appropriate responses are known to affect perceived social connection^[158].

Graceful degradation. A practical deployment principle is that Self++ should degrade gracefully when sensing or computation falls below required thresholds, rather than failing silently or maintaining a false appearance of full capability. This aligns with established guidance that automation should behave as a reliable team player by making its own limitations visible rather than masking them^[22], and with human-centred AI arguments that systems must remain safe and controllable even under reduced operating conditions^[45,87]. Concrete degradation paths follow the overlay structure. For Overlay 1, when spatial tracking becomes unreliable (e.g., due to environmental conditions, limited field-of-view, or occlusion failures), role patterns R1–R3 should fall back from spatially anchored AR cues to screen-based or audio guidance, preserving the scaffolding logic while reducing dependence on precise registration; The broader AR literature confirms that poorly registered or mistimed cues can increase rather than reduce cognitive load^[105]. For Overlay 2, when context models are uncertain or natural language interpretation is unreliable, role patterns R4–R6 should fall back from proactive advice and agentic execution to on-demand question-answering, narrowing the delegation scope to reduce the risk of out-of-the-loop failures^[144,145]; Intermediate levels of automation are preferable to full autonomy when system confidence is low^[147]. For Overlay 3, when affect recognition or social signal sensing falls below confidence thresholds, role patterns R7–R9 should suspend automated social sensing and default to user-initiated interaction, since inaccurate social inference carries higher reputational and relational risk than no inference at all; This is consistent with the XAIR recommendation that automatic

explanations should be reserved for high-confidence, high-capacity conditions and that manual, user-triggered delivery should be the default^[80]. Across all overlays, the system should disclose the nature and reason for degradation to the user (Transparency), adjust the support envelope to match the reduced capability (Adaptivity), and offer manual override or the option to restore full interaction when conditions improve (Negotiability). Degradation events should also be logged for audit, supporting the governance contestability requirements of P8.

Minimum viable Self++ and extensibility of role patterns. The nine role patterns (R1–R9) are not a closed inventory but worked examples that illustrate the design logic of each overlay. Designers may adapt, merge, subdivide, or introduce entirely new role patterns to suit domains, populations,

or capabilities not anticipated here; what Self++ prescribes is not a fixed set of roles but the structural commitments that any role pattern must satisfy: a legible supportive intent anchored within an overlay, and T.A.N. safeguards scaled to the scope and initiative of that overlay. Similarly, not all nine role patterns need to be implemented simultaneously. A minimum viable deployment could begin with a single overlay (e.g., Overlay 1 for a training application) and add overlays, or additional role patterns within an overlay, as capabilities and evaluation evidence mature. The key requirement is that whatever subset is deployed must satisfy T.A.N. at the appropriate strength for that overlay. Partial deployment also enables staged evaluation: propositions can be tested per-overlay before assessing cross-overlay interactions (P1).

Relationship to adjacent research programmes. Two emerging lines of work address complementary aspects of the design space that Self++ occupies. Recent work on embodied AI proposes taxonomies of human–AI bodily collaboration in XR, focusing on embodiment configuration: how physical or virtual bodies are shared, distributed, or swapped between human and AI partners^[190]. Zhao *et al.*'s heads-up computing programme envisions seamless computational support delivered through wearable devices in everyday scenarios, focusing on the delivery mechanism: minimising attentional cost and maximising contextual relevance^[191]. Self++ is compatible with both but addresses a dimension that neither fully treats: interactional governance over time. A embodied agent that shares motor control with a user would operate within Overlay 1 and would still need to satisfy T.A.N. constraints, transparent about which motor actions are AI-guided, adaptive to developing competence, and negotiable in control allocation. Likewise, T.A.N. can be read as governance requirements for heads-up computing, specifying the conditions under which always-on support remains beneficial rather than dependency-inducing. Self++, therefore, contributes a developmental and normative layer that these programmes currently abstract over, while they contribute embodiment-specific and delivery-specific design parameters that Self++ does not yet specify. Integrating these perspectives, embodiment configuration, delivery mechanism, and interactional governance, is a productive direction for future work.

10. Conclusion

Self++ advances a conceptual perspective on a theory of human–AI teaming for XR that treats “help” as a coupled relationship rather than a one-way service. It starts from the premise that effective augmentation must grow the person, not quietly replace them. Grounded in basic psychological needs from Self-Determination Theory (autonomy, competence, relatedness) and the Free Energy Principle's emphasis on stability under uncertainty in perception and action, Self++ frames good assistance as support that remains contestable, adjustable, and accountable.

The framework makes this actionable by organising augmentation into three interlocking overlays: Self for sensorimotor competence support, Self+ for deliberation and choice support, and Self++ for social, identity, and long-horizon alignment. These overlays are not a maturity ladder but overlays that can be activated as the situation demands. Across them, Self++ articulates role-based patterns (rather than anthropomorphic personas) and an interactional stance that keeps intent, limits, and uncertainty legible, so users can meaningfully endorse or refuse the system's contributions.

Ultimately, Self++ is a blueprint for a symbiotic cognitive niche in the spirit of J. C. R. Licklider's vision of tight human–computer partnership and the “coupled system” perspective of Andy Clark and David Chalmers. In this niche, the human supplies purpose, values, and accountable will, while the AI supplies navigable pathways, options, and scaffolding. The future is neither automated nor purely human-led, but co-determined through interactions designed to preserve agency while extending what people can perceive, decide, and become.

Declarations

Acknowledgments

The author declares that Generative AI tools, including ChatGPT 5.5 (OpenAI), Claude Opus 4.6 (Anthropic), and Gemini 3.1 (Google), as well as earlier contemporaneous versions of these tools available from early 2025 onward, were used exclusively for language editing and manuscript polishing during preparation of this work. No AI tools were used for development, analysis, interpretation of results, or formulation of conclusions.

Selected illustration components were generated using NanoBanana Pro based on author-defined prompts, concepts, and stylistic specifications. Final figure composition, integration, and editing were performed manually by the author using Photopea software.

I am deeply grateful to my former supervisor and mentor, Professor Mark Billinghurst, whose guidance shaped my path from augmented reality to empathic computing. His conviction that technology should serve people and his selflessness in supporting peers and students alike continue to inspire my work and this article.

I thank my research colleagues and students whose dedicated empirical work underpins many of the studies presented here. Self++ is, in large part, a perspective drawn from observing and reflecting on what they built; their contributions provided the evidential foundation and the motivation to write this article.

I also thank Dr Seyeon Lee for insightful feedback on the manuscript, particularly regarding the directionality of adaptivity, the dual pathways of mastery, and the generative cycling between overlays.

Authors contribution

The author contributed solely to the article.

Conflicts of interest

Thammathip Piumsomboon is an Editorial Board member of *Empathic Computing*. The authors declare that there are no other conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Funding

None.

Copyright

© The Author(s) 2026.

References

1. Licklider JC. Man-computer symbiosis. *IRE Trans Hum Factors Electron*. 1960;1:4-11. [\[DOI\]](#)
2. Engelbart DC. Augmenting human intellect: A conceptual framework (1962). In: *Ideas that created the future*. Cambridge: The MIT Press; 2021. p. 225-236. [\[DOI\]](#)
3. Clark A, Chalmers D. The extended mind. *Analysis*. 1998;58(1):7-19. [\[DOI\]](#)
4. Hutchins E. *Cognition in the wild*. Cambridge: MIT Press. 1995.
5. Sparrow B, Liu J, Wegner DM. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*. 2011;333(6043):776-778. [\[DOI\]](#)[\[PubMed\]](#)
6. Fisher M, Goddu MK, Keil FC. Searching for explanations: How the Internet inflates estimates of internal knowledge. *J Exp Psychol Gen*. 2015;144(3):674-687. [\[DOI\]](#)[\[PubMed\]](#)
7. Dahmani L, Bohbot VD. Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Sci Rep*. 2020;10(1):6310. [\[DOI\]](#)[\[PubMed\]](#)[\[PMC\]](#)
8. Kirsh D. Thinking with external representations. *AI Soc*. 2010;25(4):441-454. [\[DOI\]](#)
9. Hollan J, Hutchins E, Kirsh D. Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Trans Comput Hum Interact*. 2000;7(2):174-196. [\[DOI\]](#)
10. Bostrom N. *Superintelligence: Paths, dangers, strategies*. New York: Oxford University Press; 2014.
11. Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Yampolskiy RV, editor. *Artificial intelligence safety and security*. New York: Chapman and Hall/CRC; 2018. p. 57-69. [\[DOI\]](#)
12. Lee HH, Sarkar A, Tankelevitch L, Drosos I. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In: Yamashita N, Evers V, Yatani K, Ding X, Lee B, Chetty M, Toups-Dugas P, editors. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; 2025 Apr 26-May 1; Yokohama, Japan. New

- York: Association for Computing Machinery; 2025. p. 1-22. [DOI]
13. Pinker S. The cognitive niche: Coevolution of intelligence, sociality, and language. *Proc Natl Acad Sci U S A*. 2010;107(supplement_2):8993-8999. [DOI]
 14. Clark A. Natural-born cyborgs: Minds, technologies, and the future of human intelligence. New York: Oxford University Press; 2003.
 15. Clark A. Précis of Supersizing the mind: Embodiment, action, and cognitive extension (Oxford University Press, NY, 2008). *Philos Stud*. 2011;152(3):413-416. [DOI]
 16. Friston K. The free-energy principle: A unified brain theory? *Nat Rev Neurosci*. 2010;11(2):127-138. [DOI][PubMed]
 17. Thanyadit S, Punpongson P, Piumsomboon T, Pong TC. XR-LIVE: Enhancing asynchronous shared-space demonstrations with spatial-temporal assistive toolsets for effective learning in immersive virtual laboratories. *Proc ACM Hum Comput Interact*. 2022;6(CSCW1):1-23. [DOI]
 18. Zhang J, Han B, Dong Z, Wen R. Virtual triplets: A mixed modal synchronous and asynchronous collaboration with human-agent interaction in virtual reality. In: Mueller FF, Kyburz P, Williamson JR, Sas C, editors. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; 2024 May 11-16; Honolulu, USA. New York: Association for Computing Machinery; 2024. p. 1-8. [DOI]
 19. Zhang J, Han B, Dong Z, Topliss J, Yousefi M, Lee GA, et al. HAT swapping: Virtual agents as stand-ins for absent human instructors in virtual training. *IEEE Trans Vis Comput Graph*. 2025;31(11):9995-10004. [DOI][PubMed]
 20. Doudkin A, Pataranutaporn P, Maes P. From synthetic to human: The gap between AI-predicted and actual pro-environmental behavior change after chatbot persuasion. In: Sin J, Law E, Wallace J, Munteanu C, Korre D, editors. Proceedings of the 7th ACM Conference on Conversational User Interfaces; 2025 Jul 8-10; Waterloo, Canada. New York: Association for Computing Machinery; 2025. p. 1-18. [DOI]
 21. Liu AR, Pataranutaporn P, Maes P. The heterogeneous effects of AI companionship: An empirical model of chatbot usage and loneliness and a typology of user archetypes. *ACM Conf AI Ethics Soc*. 2025;8(2):1585-1597. [DOI]
 22. Klien G, Woods DD, Bradshaw JM, Hoffman RR, Feltoch PJ. Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intell Syst*. 2004;19(6):91-95. [DOI]
 23. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat Hum Behav*. 2024;8(12):2293-2303. [DOI]
 24. Yousefi M, Shahi A, Sharifi M, J Jorge Romera A, Hoermann S, Piumsomboon T. Team dynamics in human-AI collaboration: Effects on confidence, satisfaction, and accountability. In: Subramanian R, Nakano YI, Gedeon T, Kankanhalli M, Guha T, Shukla J, Mohammadi G, Celiktutan O, editors. Proceedings of the 27th International Conference on Multimodal Interaction; 2025 Oct 13-17; Canberra, Australia. New York: Association for Computing Machinery; 2025. p. 398-404. [DOI]
 25. Bansal G, Nushi B, Kamar E, Horvitz E, Weld DS. Is the most accurate AI the best teammate? Optimizing AI for teamwork. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2021 Feb 2-9; Singapore. Washington: Association for the Advancement of Artificial Intelligence; 2021. p. 11405-11414. [DOI]
 26. Mueller F, Semertzidis N, Andres J, Marshall J, Benford S, Li X, et al. Toward understanding the design of intertwined human-computer integrations. *ACM Trans Comput-Hum Interact*. 2023;30(5):1-45. [DOI]
 27. Zhou F, Duh HB, Billingham M. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In: Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality; 2008 Sep 15-18. Washington: IEEE Computer Society; 2008. p. 193-202. [DOI]
 28. Kim K, Billingham M, Bruder G, Duh HB, Welch GF. Revisiting trends in augmented reality research: A review of the 2nd decade of ISMAR (2008-2017). *IEEE Trans Vis Comput Graph*. 2018;24(11):2947-2962. [DOI][PubMed]
 29. Norouzi N, Kim K, Bruder G, Bailenson JN, Wisniewski P, Welch GF. The advantages of virtual dogs over virtual people: Using augmented reality to provide social support in stressful situations. *Int J Hum Comput Stud*. 2022;165:102838. [DOI]
 30. Pan X, Hamilton AFC. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *Br J Psychol*. 2018;109(3):395-417. [DOI][PubMed][PMC]
 31. Wen R, Li Q, Pu W, Mu R, Nassani A, Hoermann S, et al. GenLinguaScape: Enabling user-defined VR scenarios for communicative language practice. In: 2025 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct); 2025 Oct 8-12; Daejeon, Korea. Piscataway: IEEE; 2025. p. 831-832. [DOI]
 32. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol*. 2000;55(1):68-78. [DOI]
 33. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci*. 2013;36(3):181-204. [DOI]
 34. Vygotsky LS. Mind in society: Development of higher psychological processes. Cambridge: Harvard University Press; 1980. [DOI]
 35. Gallagher S, Raffone A, Berkovich-Ohana A, Barendregt HP, Bauer PR, Brown KW, et al. The self-pattern and Buddhist psychology. *Mindfulness*. 2024;15(4):795-803. [DOI]
 36. Wilson D, Moloney E, Parr JM, Aspinall C, Slark J. Creating an Indigenous Māori-centred model of relational health: A literature review of Māori models of health. *J Clin Nurs*. 2021;30(23-24):3539-3555. [DOI][PubMed][PMC]
 37. Markus H, Kitayama S. Culture and the self: Implications for cognition, emotion, and motivation. *Psychol Rev*. 1991;98(2):224-253. [DOI]
 38. Varela FJ, Thompson E, Rosch E. The embodied mind, revised edition: Cognitive science and human experience. Cambridge: MIT Press; 2017. [DOI]
 39. Gallagher S. How the body shapes the mind. New York: Oxford University Press; 2005. [DOI]

40. Di Paolo EA, Rohde M, De Jaegher H. Horizons for the enactive mind: Values, social interaction, and play. In: Stewart J, Gapenne O, Di Paolo EA, editors. *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge: The MIT Press; 2010. p. 32-87. [DOI]
41. Hohwy J. *The predictive mind*. New York: Oxford University Press; 2013.
42. Ho SS, Nakamura Y, Gopang M, Swain JE. Intersubjectivity as an antidote to stress: Using dyadic active inference model of intersubjectivity to predict the efficacy of parenting interventions in reducing stress: Through the lens of dependent origination in Buddhist Madhyamaka philosophy. *Front Psychol*. 2022;13:806755. [DOI]
43. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G. Active inference: A process theory. *Neural Comput*. 2017;29(1):1-49. [DOI] [PubMed]
44. Parr T, Pezzulo G, Friston KJ. *Active inference: The free energy principle in mind, brain, and behavior*. Cambridge: MIT Press; 2022.
45. Shneiderman B. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Trans Interact Intell Syst*. 2020;10(4):1-31. [DOI]
46. Capel T, Brereton M. What is human-centered about human-centered AI? A map of the research landscape. In: Schmidt A, Väänänen K, Goyal T, Kristensson O, Peters A, Mueller S, Williamson JR, Wilson ML, editors. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*; 2023 Apr 23-28; Hamburg, Germany. New York. New York: Association for Computing Machinery; 2023. p. 1-23. [DOI]
47. Amershi S, Weld D, Vorvoreanu M, Fournay A. Guidelines for human-AI interaction. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; 2019 May 4-9; Glasgow, UK. New York: Association for Computing Machinery; 2019. p. 1-13. [DOI]
48. Noggle R. The ethics of manipulation. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*. Stanford: Stanford University; 2022.
49. Faden R, Beauchamp T, King N. *A history and theory of informed consent*. New York: Oxford University Press; 1986.
50. Raz J. *The Morality of Freedom*. New York: Oxford University Press; 1988.
51. Susser D, Roessler B, Nissenbaum H. Technology, autonomy, and manipulation. *Internet Policy Rev*. 2019;8(2):1-22. [DOI]
52. Rendon-Cardona C, Burcklen MA, Legras R, Sandor C. Augmented vision systems: Paradigms and applications. *IEEE Trans Visual Comput Graphics*. 2025;31(10):9484-9501. [DOI]
53. Mori S, Ikeda S, Saito H. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Trans Comput Vis Appl*. 2017;9(1):17. [DOI]
54. Wienrich C, Latoschik ME. eXtended artificial intelligence: New prospects of human-AI interaction research. *Front Virtual Real*. 2021;2:686783. [DOI]
55. Zollmann S, Langlotz T, Grasset R, Lo WH, Mori S, Regenbrecht H. Visualization techniques in augmented reality: A taxonomy, methods and patterns. *IEEE Trans Visual Comput Graphics*. 2021;27(9):3808-3825. [DOI]
56. Dong Z, Han B, Zhang J, Wen R. An exploratory study on AI-driven visualisation techniques on decision making in extended reality. In: Viller S, Paay J, Fredericks J, Turner J, Vickery N, Wadley G, Muñoz D, Capel T, Atiq A, Davis P, Bodén M, Hardman P, Ploderer B, editors. *Proceedings of the 36th Australasian Conference on Human-Computer Interaction*; 2024 Nov 30-Dec 4; Brisbane, Australia. New York: Association for Computing Machinery; 2025. p. 654-664. [DOI]
57. Piumsomboon T, Lee GA, Hart JD, Ens B. Mini-me: An adaptive avatar for mixed reality remote collaboration. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*; 2018 Apr 21-26; Montreal, Canada. New York: Association for Computing Machinery; 2018. p. 1-13. [DOI]
58. Piumsomboon T, Lee GA, Ens B, Thomas BH, Billingham M. Superman vs giant: A study on spatial perception for a multi-scale mixed reality flying telepresence interface. *IEEE Trans Vis Comput Graph*. 2018;24(11):2974-2982. [DOI][PubMed]
59. Piumsomboon T, Lee GA, Irlitti A, Ens B, Thomas BH, Billingham M. On the shoulder of the giant: A multi-scale mixed reality collaboration with 360 video sharing and tangible interaction. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; 2019 May 4-9; Scotland, UK. New York: Association for Computing Machinery; 2019. p. 1-17. [DOI]
60. Katins C, Strecker J, Hinrichs J, Knierim P, Pfleging B, Kosch T. Ad-blocked reality: Evaluating user perceptions of content blocking concepts using extended reality. In: Yamashita N, Evers V, Yatani K, Ding X, Lee B, Chetty M, Touns-Dugas P, editors. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; 2025 Apr 26-May 1; Yokohama, Japan. New York: Association for Computing Machinery; 2025. p. 1-18. [DOI]
61. Rizzo A, Hartholt A, Grimani M, Leeds A, Liewer M. Virtual reality exposure therapy for combat-related posttraumatic stress disorder. *Computer*. 2014;47(7):31-37. [DOI]
62. Wiese W. Conscious perception as augmented reality. *Soc Epistemology*. 2026;40(1):45-58. [DOI]
63. Livingston MA, Rosenblum LJ, Brown DG, Schmidt GS, Julier SJ, Baillet Y, et al. Military applications of augmented reality. In: Furht B, editor. *Handbook of Augmented Reality*. New York: Springer; 2011. p. 671-706. [DOI]
64. Sielhorst T, Feuerstein M, Navab N. Advanced medical displays: A literature review of augmented reality. *J Display Technol*. 2008;4(4):451-467. [DOI]
65. Bonnail E, Tseng WJ, McGill M, Lecolinet E, Huron S, Gugenheimer J. Memory manipulations in extended reality. In: Schmidt A, Väänänen K, Goyal T, Kristensson PO, Peters A, Mueller S, Williamson JR, Wilson ML, editors. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*; 2023 Apr 23-28; Hamburg, Germany. New York: Association for Computing Machinery; 2023. p. 1-20. [DOI]
66. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philos Trans R Soc Lond B Biol*

- Sci. 2009;364(1535):3549-3557. [DOI][PubMed][PMC]
67. Slater M, Banakou D, Beacco A, Gallego J, Macia-Varela F, Oliva R. A separate reality: An update on place illusion and plausibility in virtual reality. *Front Virtual Real.* 2022;3:914392. [DOI]
 68. Triberti S, Sapone C, Riva G. Being there but where? Sense of presence theory for virtual reality applications. *Humanit Soc Sci Commun.* 2025;12:79. [DOI]
 69. Skarbez R, Brooks FP, Whitton MC. Immersion and coherence: Research agenda and early results. *IEEE Trans Vis Comput Graph.* 2021;27(10):3839-3850. [DOI][PubMed]
 70. Rastelli C, Greco A, Kenett YN, Finocchiaro C, De Pisapia N. Simulated visual hallucinations in virtual reality enhance cognitive flexibility. *Sci Rep.* 2022;12:4027. [DOI]
 71. Job M, Manoni M, Sansone LG, Viceconti A, Testa M. A surprise induced by a visual-haptic illusion in virtual reality can lead to motor improvement. *Sci Rep.* 2025;15:14741. [DOI]
 72. Skinner BF. *The behavior of organisms: An experimental analysis.* New York: Appleton-Century-Crofts; 2019.
 73. Piumsomboon T, Ong G, Urban C, Ens B, Topliss J, Bai X, et al. Ex-Cit XR: Expert-elicitation and validation of Extended Reality visualisation and interaction techniques for disengaging and transitioning users from immersive virtual environments. *Front Virtual Real.* 2022;3:943696. [DOI]
 74. Yang X, Sasikumar P, Amtsberg F, Menges A, Sedlmair M, Nanayakkara S. Who is in control? Understanding user agency in AR-assisted construction assembly. In: Yamashita N, Evers V, Yatani K, Ding X, Lee B, Chetty M, Touns-Dugas P, editors. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; 2025 April 26-May 1; Yokohama, Japan. New York: Association for Computing Machinery; 2025. p. 1-15. [DOI]
 75. Seeber I, Bittner E, Briggs RO, de Vreede T, de Vreede GJ, Elkins A, et al. Machines as teammates: A research agenda on AI in team collaboration. *Inf Manag.* 2020;57(2):103174. [DOI]
 76. Zhang R, McNeese NJ, Freeman G, Musick G. "An ideal human": Expectations of AI teammates in human-AI teaming. *Proc ACM Hum-Comput Interact.* 2021;4(CSCW3):1-25. [DOI]
 77. Norman DA. *The psychology of everyday things.* New York: Basic Books, Inc.; 1988.
 78. Duan W, Flathmann C, McNeese N, Scalia MJ. Trusting autonomous teammates in human-AI teams - a literature review. In: Yamashita N, Evers V, Yatani K, Ding X, Lee B, Chetty M, Touns-Dugas P, editors. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; 2025 Apr 26-May 1; Yokohama, Japan. New York: Association for Computing Machinery; 2025. p. 1-23. [DOI]
 79. Han B, Dong Z, Zhang J, Wen R, Hirai T, Clark A, et al. Exploring mediation by an embodied virtual agent in immersive triadic collaborative decision-making. *IEEE Trans Vis Comput Graph.* 2026. [DOI][PubMed]
 80. Xu X, Yu A, Jonker TR, Todi K. XAIR: A framework of explainable AI in augmented reality. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* ACM, 2023:1-30. [DOI]
 81. Nam H, Kang S, Woo W, Kim K. AVAGENT: Bridging asynchronous communication through AI-powered virtual avatars. In: *2025 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*; 2025 Mar 8-12; Saint Malo, France. Piscataway: IEEE; 2025. p. 1142-1146. [DOI]
 82. Yang J, Tan R, Wu Q, Zheng R, Peng B, Liang Y, et al. Magma: A foundation model for multimodal AI agents. In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2025 Jun 10-17; Nashville, USA. Piscataway: IEEE; 2025. p. 14203-14214. [DOI]
 83. Liao QV, Vaughan JW. AI transparency in the age of LLMs: A human-centered research roadmap. *arXiv:2306.01941 [Preprint]*. 2023. [DOI]
 84. Li C, Wu G, Chan GY, Turakhia DG. Satori: Towards proactive AR assistant with belief-desire-intention user modeling. In: Yamashita N, Evers V, Yatani K, Ding X, Lee B, Chetty M, Touns-Dugas P, editors. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; 2025 Apr 26-May 1; Yokohama, Japan. New York: Association for Computing Machinery; 2025. p. 1-24. [DOI]
 85. Lee M, Liang P, Yang Q. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In: Barbosa S, Lampe C, Appert C, Shamma DA, Drucker S, Williamson J, Yatani K, editors. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*; 2022 Apr 29-May 5; New Orleans, USA. New York: Association for Computing Machinery; 2022. p. 1-19. [DOI]
 86. Nishal S, Lee M, Diakopoulos N, Wortman Vaughan J. "Helping me versus doing it for me": Designing for agency in LLM-infused writing tools for science journalism. In: Oliver N, Shamma DA, Candello H, Cesar P, Lopes P, Bozzon A, Kosch T, Liao V, Ma X, Artizzu V, Draxler F, López G, Reinschluessel AV, Tong X, Touns Dugas PO, editors. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*; 2026 Apr 13-17; Barcelona, Spain. New York: Association for Computing Machinery; 2026. p. 1-20. [DOI]
 87. Shneiderman B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int J Hum.* 2020;36(6):495-504. [DOI]
 88. Yang Q, Steinfeld A, Rosé C, Zimmerman J. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*; 2020 Apr 25-30; Honolulu, USA. New York: Association for Computing Machinery; 2020. p. 1-13. [DOI]
 89. Yousefi M, Crowe SE, Hoermann S, Sharifi M, Romera A, Shahi A, et al. Advancing prosociality in extended reality: Systematic review of the use of embodied virtual agents to trigger prosocial behaviour in extended reality. *Front Virtual Real.* 2024;5:1386460. [DOI]
 90. Kim K, Boelling L, Haesler S, Bailenson J, Bruder G, Welch GF. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In: *2018 IEEE International Symposium on Mixed and*

- Augmented Reality (ISMAR); 2018 Oct 16-20; Munich, Germany. Piscataway: IEEE; 2018. p. 105-114. [DOI]
91. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: Interactive simulacra of human behavior. In: Follmer S, Han J, Steimle J, Riche NH, editors. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology; 2023 Oct 29-Nov 1; San Francisco, USA. New York: Association for Computing Machinery; 2023. p. 1-22. [DOI]
 92. Behrouz A, Razaviyayn M, Zhong P, Mirrokni V. Nested learning: The illusion of deep learning architectures. arXiv:2512.24695 [Preprint]. 2025. [DOI]
 93. Horvitz E. Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems; 1999 May 15-20; Pittsburgh, USA. New York: Association for Computing Machinery; 1999. p. 159-166. [DOI]
 94. Bradshaw JM, Sierhuis M, Acquisti A, Feltovich P, Hoffman R, Jeffers R, et al. Adjustable autonomy and human-agent teamwork in practice: An interim report on space applications. In: Hexmoor H, Castelfranchi C, Falcone R, editors. Agent autonomy. Boston: Springer; 2003. p. 243-280. [DOI]
 95. Greunke L, Sadagic A. Taking immersive VR leap in training of landing signal officers. IEEE Trans Vis Comput Graph. 2016;22(4):1482-1491. [DOI][PubMed]
 96. Orlosky J, Sra M, Bektaş K, Peng H, Kim J, Kos'myna N, et al. Telelife: The future of remote living. Front Virtual Real. 2021;2:763340. [DOI]
 97. Jing A, May K, Lee G, Billingham M. Eye see what you see: Exploring how bi-directional augmented reality gaze visualisation influences co-located symmetric collaboration. Front Virtual Real. 2021;2:697367. [DOI]
 98. Turkle S. Alone together: Why we expect more from technology and less from each other. New York: Basic Books; 2011.
 99. Dreyfus SE. The five-stage model of adult skill acquisition. Bull Sci Technol Soc. 2004;24(3):177-181. [DOI]
 100. Lee GA, Teo T, Kim S, Billingham M. A user study on MR remote collaboration using live 360 video. In: 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR); 2018 Oct 16-20; Munich, Germany. Piscataway: IEEE; 2018. p. 153-164. [DOI]
 101. Oda O, Elvezio C, Sukan M, Feiner S, Tversky B. Virtual replicas for remote assistance in virtual and augmented reality. In: Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology; 2015 Nov 11-15; Charlotte, USA. New York: Association for Computing Machinery; 2015. p. 405-415. [DOI]
 102. Huang G, Qian X, Wang T, Patel F. AdapTutAR: An adaptive tutoring system for machine tasks in augmented reality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021 May 8-13; Yokohama, Japan. New York: Association for Computing Machinery; 2021. p. 1-15. [DOI]
 103. Anderson JR, Corbett AT, Koedinger KR, Pelletier R. Cognitive tutors: Lessons learned. J Learn Sci. 1995;4(2):167-207. [DOI]
 104. Vanneste P, Huang Y, Park JY, Cornillie F, Decloedt B, Van den Noortgate W. Cognitive support for assembly operations by means of augmented reality: An exploratory study. Int J Hum Comput Stud. 2020;143:102480. [DOI]
 105. Buchner J, Buntins K, Kerres M. The impact of augmented reality on cognitive load and performance: A systematic review. J Comput Assist Learn. 2022;38(1):285-303. [DOI]
 106. Atkinson RK, Maier UH. From studying examples to solving problems: Fading worked-out solution steps helps learning. In: Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society; 2000 Aug 13-15; Philadelphia: University of Pennsylvania. UK: Psychology Press; 2000. Available from: <https://escholarship.org/uc/item/81b9j9hs>
 107. Sweller J, Ayres P, Kalyuga S. The guidance fading effect. In: Cognitive load theory. New York: Springer; 2011. p. 171-182. [DOI]
 108. Schmidt RA. A schema theory of discrete motor skill learning. Psychol Rev. 1975;82(4):225-260. [DOI]
 109. Raviv L, Lupyan G, Green SC. How variability shapes learning and generalization. Trends Cogn Sci. 2022;26(6):462-483. [DOI]
 110. Domínguez A, Pacho G, Bowers L, Wild F, Alcock S, Chiazese G, et al. Dataset of user interactions across four large pilots on the use of augmented reality in learning experiences. Sci Data. 2023;10(1):823. [DOI][PubMed][PMC]
 111. Cho H, Chang E, Yuan B, Teo T, Lee GA, Piumsomboon T, et al. Bichronous collaboration: Using spatiotemporal cues to collaborate across time and space on physical tasks. In: 2025 IEEE international symposium on mixed and augmented reality (ISMAR); 2025 Oct 8-12; Daejeon, Korea. Piscataway: IEEE; 2025. p. 1398-1408. [DOI]
 112. Yang U, Kim GJ. Implementation and evaluation of "just follow me": An immersive, VR-based, motion-training system. Presence Teleoperators Virtual Environ. 2002;11(3):304-323. [DOI]
 113. Jarc AM, Stanley AA, Clifford T, Gill IS, Hung AJ. Proctors exploit three-dimensional ghost tools during clinical-like training scenarios: A preliminary study. World J Urol. 2017;35(6):957-965. [DOI]
 114. Piumsomboon T, Altamira D, Kim H, Clark A, Lee G, Billingham M. Grasp-Shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In: 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR); 2014 Sep 10-12; Munich, Germany. Piscataway: IEEE; 2014. p. 73-82. [DOI]
 115. Limbu BH, Jarodzka H, Klemke R, Specht M. Using sensors and augmented reality to train apprentices using recorded expert performance: A systematic literature review. Educ Res Rev. 2018;25:1-22. [DOI]
 116. Kirschner PA, Sweller J, Kirschner F, Zambrano R J. From cognitive load theory to collaborative cognitive load theory. Intern J Comput-Support Collab Learn. 2018;13(2):213-233. [DOI]
 117. Renkl A. The worked examples principle in multimedia learning. In: The Cambridge handbook of multimedia learning. Cambridge: Cambridge University Press; 2014. p. 391-412. [DOI]
 118. Collins A, Brown JS, Newman SE. Cognitive apprenticeship: Teaching the Crafts of reading, writing, and mathematics. In: Resnick LB, editor. Knowing, learning, and instruction. Hillsdale: Lawrence Erlbaum Associates; 2018. p. 453-494. [DOI]
 119. Csikszentmihalyi, Mihaly.(1990) . Flow: The psychology of optimal experience. J Leis Res. 1992;24(1):93-94. [DOI]

120. Campero A, Raileanu R, Küttler H, Tenenbaum JB, Rocktäschel T, Grefenstette E. Learning with AMIGo: Adversarially motivated intrinsic goals. *arXiv:2006.12122 [Preprint]*. 2020. [DOI]
121. Bjork EL, Bjork RA. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In: Gernsbacher MA, Pomerantz J, editors. *Psychology and the real world: Essays illustrating fundamental contributions to society*; New York: Worth Publishing; 2014. p. 59-68. Available from: <https://jacobzelko.com/05252020211350-hard-on-self/>
122. Landman A, van Oorschot P, René van Paassen MMR, Groen EL, Bronkhorst AW, Mulder M. Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios. *Hum Factors*. 2018;60(6):793-805. [DOI][PubMed][PMC]
123. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychol Rev*. 1993;100(3):363-406. [DOI]
124. Sarter NB, Woods DD. How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Hum Factors*. 1995;37(1):5-19. [DOI]
125. Eom H, Lee SH. Mode confusion of human-machine interfaces for automated vehicles. *J Comput Des Eng*. 2022;9(5):1995-2009. [DOI]
126. Lyons JB, Sycara K, Lewis M, Capiola A. Human-autonomy teaming: Definitions, debates, and directions. *Front Psychol*. 2021;12:589585. [DOI]
127. Wischnewski M, Krämer N, Müller E. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In: Schmidt A, Väänänen K, Goyal T, Kristensson PO, Peters A, Mueller S, Williamson JR, Wilson ML, editors. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*; Hamburg, Germany. New York: Association for Computing Machinery; 2023. p. 1-16. [DOI]
128. Okamura K, Yamada S. Adaptive trust calibration for human-AI collaboration. *PLoS One*. 2020;15(2):e0229132. [DOI]
129. Beer JM, Fisk AD, Rogers WA. Toward a framework for levels of robot autonomy in human-robot interaction. *J Hum Robot Interact*. 2014;3(2):74-99. [DOI][PubMed][PMC]
130. Doudkin A, Pataranutaporn P, Maes P. AI persuading AI vs AI persuading humans: LLMs' differential effectiveness in promoting pro-environmental behavior. *arXiv:2503.02067 [Preprint]*. 2025. [DOI]
131. Mathur A, Acar G, Friedman MJ, Lucherini E, Mayer J, Chetty M, et al. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proc ACM Hum-Comput Interact*. 2019;3:1-32. [DOI]
132. Luguri J, Strahilevitz LJ. Shining a light on dark patterns. *J Leg Anal*. 2021;13(1):43-109. [DOI]
133. Thaler RH, Sunstein CR. *Nudge: Improving decisions about health, wealth, and happiness*. New Haven: Yale University Press. 2008.
134. Sunstein CR. Nudging and choice architecture: Ethical considerations. *Yale J Regul*. 2015. Available from: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:17915544>
135. Schmidt AT, Engelen B. The ethics of nudging: An overview. *Philos Compass*. 2020;15(4):e12658. [DOI]
136. Tonniss M, Klein L, Klinker G. Perception thresholds for augmented reality navigation schemes in large distances. In: 2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality; 2008 Sep 15-18; Cambridge, UK. Piscataway: IEEE; 2008. p. 189-190. [DOI]
137. Kim S, Dey AK. Simulated augmented reality windshield display as a cognitive mapping aid for elder driver navigation. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; 2009 Apr 4-9; Boston, USA. New York: Association for Computing Machinery; 2009. p. 133-142. [DOI]
138. Meske C, Amojó I. Ethical guidelines for the construction of digital nudges. *arXiv:2003.05249v1 [Preprint]*. 2020. [DOI]
139. Sorensen T, Moore J, Fisher J, Gordon M, Mireshghallah N, Rytting CM, et al. A roadmap to pluralistic alignment. *arXiv:2402.05070 [Preprint]*. 2024. [DOI]
140. Reicherts L, Zhang ZT, von Oswald E, Liu Y, Rogers Y, Hassib M. AI, help me think: But for myself: Assisting people in complex decision-making by providing different kinds of cognitive support. In: Yamashita N, Evers V, Yatani K, Ding X, Lee B, Chetty M, Touns-Dugas P, editors. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; 2025 Apr 26-May 1; Yokohama, Japan. New York: Association for Computing Machinery; 2025. p. 1-19. [DOI]
141. Haque AB, Islam AKMN, Mikalef P. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technol Forecast Soc Change*. 2023;186:122120. [DOI]
142. Steyvers M, Kumar A. Three challenges for AI-assisted decision-making. *Perspect Psychol Sci*. 2024;19(5):722-734. [DOI]
143. Krakowski S. Human-AI agency in the age of generative AI. *Inf Organ*. 2025;35(1):100560. [DOI]
144. Endsley MR. From here to autonomy: Lessons learned from human-automation research. *Hum Factors*. 2017;59(1):5-27. [DOI]
145. Endsley MR, Kiris EO. The out-of-the-loop performance problem and level of control in automation. *Hum Factors*. 1995;37(2):381-394. [DOI]
146. Cheng EC, Cheng J, Siu A. Toward safe and responsible AI agents: A three-pillar model for transparency, accountability, and trustworthiness. *arXiv:2601.06223 [Preprint]*. 2026. [DOI]
147. Kaber DB, Endsley MR. Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. *Process Saf Prog*. 1997;16(3):126-131. [DOI]
148. Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk B, Sierhuis M. The fundamental principle of coactive design: Interdependence must shape autonomy. In: De Vos M, Fornara N, Pitt JV, Vouras G, editors. *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*. Berlin: Springer; 2011. p. 172-191. [DOI]
149. Mathieu JE, Heffner TS, Goodwin GF, Salas E, Cannon-Bowers JA. The influence of shared mental models on team process and performance. *J Appl Psychol*. 2000;85(2):273-283. [DOI]

150. De Dreu CKW, Weingart LR. Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *J Appl Psychol.* 2003;88(4):741-749. [DOI]
151. DeChurch LA, Mesmer-Magnus JR. The cognitive underpinnings of effective teamwork: A meta-analysis. *J Appl Psychol.* 2010;95(1):32-53. [DOI][PubMed]
152. Weick KE. Sensemaking in organizations. Thousand Oaks: SAGE Publications, Inc; 1995.
153. Chen H, Wang P, Hao S. AI in the spotlight: The impact of artificial intelligence disclosure on user engagement in short-form videos. *Comput Hum Behav.* 2025;162:108448. [DOI]
154. Lukosch S, Billinghamurst M, Alem L, Kiyokawa K. Collaboration in augmented reality. *Comput Supported Coop Work.* 2015;24(6):515-525. [DOI]
155. Piumsomboon T, Dey A, Ens B, Lee G, Billinghamurst M. The effects of sharing awareness cues in collaborative mixed reality. *Front Robot AI.* 2019;6:5. [DOI]
156. Kim S, Lee G, Huang W, Kim H, Woo W, Billinghamurst M. Evaluating the combination of visual communication cues for HMD-based mixed reality remote collaboration. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; 2019 May 4-9; Glasgow, UK. New York: Association for Computing Machinery; 2019. p. 1-13. [DOI]
157. Kim T, Chang A, Holland L, Pentland AS. Meeting mediator: Enhancing group collaboration using sociometric feedback. In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work*; 2008 Nov 8-12; San Diego, USA. New York: Association for Computing Machinery; 2008. p. 457-466. [DOI]
158. Templeton EM, Chang LJ, Reynolds EA, Cone LeBeaumont MD, Wheatley T. Fast response times signal social connection in conversation. *Proc Natl Acad Sci U S A.* 2022;119(4):e2116915119. [DOI][PubMed][PMC]
159. De Freitas J, Oğuz-Uğuralp Z, Uğuralp AK, Puntoni S. AI companions reduce loneliness. *J Consum Res.* 2026;52(6):1126-1148. [DOI]
160. Deci EL, Ryan RM. Self-determination theory. In: Lange PAMV, Kruglanski AW, Higgins ET, editors. *Handbook of Theories of Social Psychology*. Thousand Oaks: SAGE Publications; 2012. p. 416-436. [DOI]
161. Peck TC, Seinfeld S, Aglioti SM, Slater M. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Conscious Cogn.* 2013;22(3):779-787. [DOI][PubMed]
162. Slater M, Gonzalez-Liencre C, Haggard P, Vinkers C, Gregory-Clarke R, Jelley S, et al. The ethics of realism in virtual and augmented reality. *Front Virtual Real.* 2020;1:1. [DOI]
163. Pataranutaporn P, Winson K, Yin P, Lapapirojn A, Ouppaphan P, Lertsutthiwong M, et al. Future you: A conversation with an AI-generated future self reduces anxiety, negative emotions, and increases future self-continuity. In: *2024 IEEE Frontiers in Education Conference (FIE)*; 2024 Oct 13-16; Washington, USA. Piscataway: IEEE; 2024. p. 1-10. [DOI]
164. Hershfield HE, Goldstein DG, Sharpe WF, Fox J, Yeykelis L, Carstensen LL, et al. Increasing saving behavior through age-progressed renderings of the future self. *J Mark Res.* 2011;48:S23-S37. [DOI]
165. Chen VHH, Ibasco GC. All it takes is empathy: How virtual reality perspective-taking influences intergroup attitudes and stereotypes. *Front Psychol.* 2023;14:1265284. [DOI]
166. Markowitz DM, Laha R, Perone BP, Pea RD, Bailenson JN. Immersive virtual reality field trips facilitate learning about climate change. *Front Psychol.* 2018;9:2364. [DOI]
167. Herrmann T, Pfeiffer S. Keeping the organization in the loop: A socio-technical extension of human-centered artificial intelligence. *AI & Soc.* 2023;38(4):1523-1542. [DOI]
168. Leofante F, Ayoobi H, Dejl A, Freedman G, Gorur D, Jiang J, et al. Contestable AI needs computational argumentation. In: Marquis P, Ortiz M, Pagnucco M, editors. *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*; 2024 Nov 2-8; Hanoi, Vietnam. California: IJCAI Organization; 2024. p. 888-896. [DOI]
169. Ashton H. Causal Campbell-Goodhart's law and Reinforcement Learning. *arXiv:2011.01010 [Preprint]*. 2020. [DOI]
170. Karwowski J, Hayman O, Bai X, Kiendlhofer K, Griffin C, Skalse J. Goodhart's law in reinforcement learning. *arXiv:2310.09144 [Preprint]*. 2023. [DOI]
171. Macy JR. Dependent co-arising: The distinctiveness of Buddhist ethics. *J Relig Ethics.* 1979;7(1):38-52. Available from: <http://www.jstor.org/stable/40018242>
172. Buxton B. *Sketching user experiences: Getting the design right and the right design*. San Francisco: Morgan Kaufmann Publishers Inc.; 2010.
173. Höök K, Löwgren J. Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Trans Comput-Hum Interact.* 2012;19(3):1-18. [DOI]
174. Gaver W. What should we expect from research through design? In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; 2012 May 5-10; Austin, USA. New York: Association for Computing Machinery; 2012. p. 937-946. [DOI]
175. Dourish P. Implications for design. In: Grinter R, Rodden T, Aoki P, Cutrell E, Jeffries R, Olson G, editors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; 2006 Apr 22-27; Montréal, Canada. New York: Association for Computing Machinery; 2006. p. 541-550. [DOI]
176. Carroll JM, Rosson MB. Getting around the task-artifact cycle: How to make claims and design by scenario. *ACM Trans Inf Syst.* 1992;10(2):181-212. [DOI]
177. Gregor S, Jones D. The anatomy of a design theory. *J Assoc Inf Syst.* 2007;8(5):312-335. [DOI]
178. Lee JD, See KA. Trust in automation: Designing for appropriate reliance. *Hum Factors.* 2004;46(1):50-80. [DOI]

179. Davidson K, Lisle L, Whitley K, Bowman DA, North C. Exploring the evolution of sensemaking strategies in immersive space to think. *IEEE Trans Vis Comput Graph*. 2023;29(12):5294-5307. [DOI][PubMed]
180. Hart JD, Piumsomboon T, Lee GA, Smith RT, Billinghamurst M. Manipulating avatars for enhanced communication in extended reality. In: 2021 IEEE International Conference on Intelligent Reality (ICIR); 2021 May 12-13; Piscataway, USA. Piscataway: IEEE; 2021, p. 9-16. [DOI]
181. Kitson A, Chirico A, Gaggioli A. A review on research and evaluation methods for investigating self-transcendence. *Front Psychol*. 2020;11:547687. [DOI]
182. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550(7676):354-359. [DOI][PubMed]
183. Russell S. *Human compatible: Artificial intelligence and the problem of control*. New York: Viking; 2019.
184. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. *arXiv:1606.06565v2* [Preprint]. 2016. [DOI]
185. Burns C, Izmailov P, Kirchner JH, Baker B, Gao L, Aschenbrenner L, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv:2312.09390v1* [Preprint]. 2023. [DOI]
186. Olah C, Mordvintsev A, Schubert L. Feature visualization. *Distill*. 2017;2(11):e7. [DOI]
187. Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, et al. The building blocks of interpretability. *Distill*. 2018;3(3):e10. [DOI]
188. Christiano PF, Leike J, Brown T. Deep reinforcement learning from human preferences. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017 Dec 4-9; California, USA. United States: Curran Associates Inc.; 2017. p. 4302-4310. [DOI]
189. Kruger J, Dunning D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol*. 1999;77(6):1121-1134. [DOI]
190. Lu F, Zhao Q. Towards cobodied/symbodied AI: Concept and eight scientific and technical problems. *Sci China Inf Sci*. 2026;69:116101. [DOI]
191. Zhao S, Tan F, Fennedy K. Heads-up computing moving beyond the device-centered paradigm. *Commun ACM*. 2023;66(9):56-63. [DOI]

Publisher's Note

Science Exploration remains a neutral stance on jurisdictional claims in published maps and institutional affiliations. The views expressed in this article are solely those of the author(s) and do not reflect the opinions of the Editors or the publisher.