ChatGPT

Exploring User Consent Frameworks for Human-Centric Interactions

Understanding Consent Frameworks in Human-AI Interaction

Designing interactive systems with a **user consent framework** means embedding principles that ensure users maintain control over what happens in a conversation or platform. Much like *Constitutional AI* (which guides AI behavior via a set of written principles [1]), a consent framework provides rules that prioritize the user's autonomy, safety, and comfort. This is crucial because non-consensual or boundary-crossing interactions – from unwanted data sharing to online harassment – are pervasive and harmful [2]. By structuring interactions around consent, we **unlock more human-centric experiences** where users feel safe and respected. In practice, this correlates strongly with **user safety** (preventing unwanted or harmful experiences) and **user satisfaction** (users feel heard, in control, and trusting of the system). Researchers have increasingly argued that any design involving fluent, human-like AI or sensitive user data *"must center user consent"* [3] to preserve user agency and well-being.

Key Principles of a Consent Framework

Insights from ethics, law, and human-computer interaction (HCI) have led to concrete principles for consent-centric design. A notable model is the **affirmative consent framework** (originating from *"Yes means yes"* sexual consent doctrine, now adapted to tech platforms [2]). This framework defines consent with five key attributes [4] [5]:

- **Voluntary:** Consent should be freely given without coercion. The user should willingly *opt-in* to an interaction or feature, rather than being automatically included.
- **Informed:** Users must have clear, accessible information about what they are consenting to. In other words, no hidden motives or fine print – an informed user is aware of potential outcomes or data use before saying "yes" [6].
- **Revertible (Revocable):** Consent is not a one-time forever deal – it can be withdrawn at any point [7]. Systems should let users easily undo or opt out of actions (e.g. leaving a group, revoking data permissions) to respect changing minds or comfort levels.
- **Specific:** Agreement to one action is *not* blanket consent for others [8] [9]. Consent should be scoped – for example, a user may consent to share a post with friends but not with the public. The system should enable granular controls so users can consent to particular people or content types and not implicitly to all.
- **Unburdensome:** Obtaining or managing consent should not be so cumbersome that users feel compelled to agree just to avoid hassle [10] [11]. If saying "no" is difficult or if users are fatigued by constant permission prompts, consent isn't truly free. A consentful system minimizes repetitive prompts and makes privacy settings easy, so that *saying no is as easy as saying yes* [12].

These principles have been proposed as a foundation for **consentful technology** on social platforms [13] [4]. They aim to protect user boundaries and agency online. For instance, researchers Im et al. applied

affirmative consent as a lens to re-imagine social media, addressing issues like unwanted tagging, harassment, or content exposure [14] [15]. By prioritizing *voluntary, revocable, and specific* user decisions, platforms can curb non-consensual harms (e.g. strangers auto-adding you to groups or malicious actors exploiting your data) and thus improve safety. Notably, affirmative consent inherently emphasizes user **agency** – *the user "owns" their interactions and data* and can dictate what happens with them [16]. This focus on agency is especially empowering for marginalized or vulnerable users who historically have had their consent and boundaries overridden [16]. In short, a consent framework built on these principles makes digital interactions more equitable and user-centered.

Consent in Emotionally Charged Situations

Consent becomes even more complex – and crucial – in **emotionally charged or high-stakes conversations**. When strong feelings (anger, fear, distress, etc.) are involved, users may be more vulnerable or prone to harm, and misunderstandings can escalate quickly. Ensuring consent in these contexts means checking that all parties are emotionally ready and willing to continue, and that they agree on how to proceed. In conflict resolution and counseling, this is a well-known practice: for example, practitioners of *Nonviolent Communication (NVC)* always *"ask permission"* before using its techniques with someone new, explicitly giving the person the chance to pause or stop if they feel uncomfortable [17] [18]. As one trainer put it, *"real permission, giving consent… builds trust… honesty and transparency… It's worth slowing down for."* [18] This highlights that pausing to obtain consent – even mid-conversation – can foster trust, because the user sees that their emotional state and boundaries matter more than just pushing the conversation forward.

Emotional situations also test the **limits of 'informed' and 'voluntary' consent**. A user who is upset or afraid might agree to something in the moment that they wouldn't under calmer circumstances. Psychology research suggests that emotion and empathy do influence decision-making in consent; for instance, individuals with greater emotional awareness tend to show *more caution in giving consent* (higher rates of refusal when something doesn't feel right) [19]. This implies that a distressed user might lack full decisional capacity – they could either agree too easily due to impaired focus, or conversely refuse everything out of fear. A consent framework should account for this by perhaps **slowing down interactions** in heated moments and double-checking important decisions. Some emerging design ideas include AI systems that detect emotional cues and then **ask for explicit confirmation** or offer to pause the discussion when emotions run high. Such measures ensure that consent is genuine and not given under duress or confusion.

In multi-party conflicts or potentially harmful disputes, **consent must extend to the process itself**. All participants should agree on how the discussion or mediation will be conducted and that they wish to continue engaging. If one party feels overwhelmed or trapped, the interaction ceases to be truly consensual. Traditional dispute resolution frameworks distinguish *consensual processes* (like mediation or negotiation, where parties retain control of the outcome) from adjudicatory ones (like court trials, where control is handed to a judge) [20]. Outcomes from consensual processes tend to be more satisfactory because everyone's agreement is required for a resolution [20]. Translating this to AI-mediated chat or online forums: features can allow any user to *"opt out"* or call a timeout in a heated discussion. By making an **easy exit or pause button** available, users know they are not forced to endure an interaction that feels unsafe. This ability to withdraw (the *revertible* principle) on-the-fly is critical in emotional scenarios. It not only protects users, but also incentivizes more respectful behavior – participants know the conversation continues only by mutual consent.

Designing Consentful Interactions Without Losing Flow

One of the toughest challenges is implementing consent mechanisms that are **trackable and robust (for accountability)** yet **subtle enough not to disrupt conversation flow**. To be effective as evidence, consent actions need to be logged in a clear, tamper-proof way – especially if they might later be needed in court to prove what a user agreed to. At the same time, constantly interrupting a chat with formal consent prompts ("Do you agree? [Yes/No]") can frustrate users and hinder natural engagement. Researchers refer to this tension as the *consent fatigue* problem: when asked to micromanage consent too often, users get annoyed or simply acquiesce without true consideration [21]. Thus, the key is finding a balance where **user consent is continuously respected and recorded, but not endlessly demanded from the user**.

Several approaches are being explored to achieve this balance:

- **Periodic and Contextual Check-Ins:** Instead of nagging users at every step, systems can intelligently time their consent prompts. For example, a social platform might **periodically ask if a user wants to remain in a group chat** they were added to, rather than auto-adding them without choice [22] [23]. Contextual triggers can also help; an AI could issue a gentle prompt like, "*This next section might be sensitive. Do you want to continue?*" when a conversation is about to enter potentially harmful territory. By **front-loading warnings and requests for permission**, the interaction remains fluid—users are only interrupted when it truly matters to their safety or preferences.

- **Granular Consent Controls:** Designing the interface to let users set **preferences in advance** reduces the need for repeated prompts. Social platforms inspired by the affirmative consent model have introduced ideas like **"granular visibility"** settings and **social circles** [24] [25]. Users could predefine who can contact them, what topics they are comfortable with, and how far their content can spread (e.g. share with friends-of-friends vs. public) [26] [27]. The system then respects these consent settings automatically. If a boundary is about to be crossed, the system can either block the action or ask for an override consent just for that instance. Because the user's choices are *remembered* and enforceable by design, it feels less like a constant interrogation and more like the platform simply knowing the user's boundaries.

- **Ongoing but Unobtrusive Consent Indicators:** Another idea is to make consent **"ambient"** in the UI – visible and adjustable without requiring a full stop. For instance, in a video call or VR setting, an icon could glow green when all participants' consent conditions are satisfied and turn yellow/red if something is amiss (someone's recording without permission, or a topic entered a no-go zone). This gives a subtle signal to check in, without an abrupt alert. Participants could then address it: e.g., *"I notice you're uncomfortable – shall we change the topic?"* This kind of design is preventive and responsive: it reframes potential harm as *"unwanted experiences happening due to lack of consent mechanisms"* [28]. If those mechanisms (like clear indicators or quick opt-out buttons) are in place, many harms can be avoided by **catching consent violations early** rather than after damage is done [29] [30].

- **Secure Consent Logging:** To make consent **trackable as evidence**, researchers are experimenting with technologies like blockchain. One recent proposal coined *"demonstrated consent"* uses blockchain records and even non-fungible tokens to create an immutable log of what a user consented to and when [31]. Paired with AI (e.g. large language models to explain terms and confirm understanding), this system was aimed at biobank data consent, but the concept extends to

conversation logs as well. The idea is to provide a **secure, transparent ledger** of user consent that can be easily reviewed later [32]. In a chat scenario, this might mean every time a user grants or revokes permission (say, to share their message or to continue a sensitive dialogue), the event is timestamped on an append-only ledger. Such a record could indeed serve as reliable evidence of consent (or non-consent) in legal disputes. The challenge is to integrate this seamlessly – potentially the logging happens in the background, unless a dispute arises and the record is needed. Users could even receive a **"consent receipt"** after significant actions, summarizing what they agreed to [31]. This keeps users informed and provides transparency without requiring them to fill out lengthy forms mid-chat.

Despite these promising directions, **no perfect solution exists yet**. Each approach must overcome trade-offs. Too few interruptions and we risk ambiguity about consent; too many and we risk user frustration and disengagement [21]. The design community emphasizes iterative co-design with users to get this right: in one participatory study, users themselves suggested features like *pre-conversation consent contracts*, content safewords, or "Are we still okay?" pop-up checks in a heated VR dating scenario [33] [30]. Such user-centered ideas are invaluable for crafting mechanisms that feel natural. The ultimate goal is a **consent framework that operates like a gentle safety net** – mostly unobtrusive, but always there to catch potential violations of trust.

Conclusion and Future Outlook

Consent frameworks represent a new frontier in making AI and online interactions more **ethical, user-centric, and emotionally intelligent**. They borrow principles from domains like sexual consent law, psychology, and conflict resolution, and apply them to technology design. Existing research – from *affirmative consent* on social media [4] [5] to *consentful VR* interfaces [33] [30] – shows that empowering users to actively say *yes* or *no* at crucial moments greatly enhances safety and trust. Users feel **safer and more satisfied** when they can control their level of engagement and know that the system will back them up if they set a boundary. At the same time, implementing these ideas is complex. Emotional nuances, power imbalances, and practical UX constraints mean there's no one-size-fits-all solution. For instance, an approach that works well for consenting to data sharing (like blockchain logs) might not directly solve the problem of consenting to a sensitive personal conversation in real-time.

What is clear is that **consent in human-AI and human-human digital interactions is an active area of research**, with interdisciplinary efforts underway. Technologists are collaborating with ethicists, legal experts, and mental health professionals to refine consent mechanisms. Future systems might combine several techniques discussed: e.g. an AI assistant that *negotiates consent continually* – it might explain risks (keeping users informed), monitor emotional tone (suggest breaks or get affirmation in emotional moments), and transparently log agreements (for accountability). The vision is to create digital environments that are as respectful of personal boundaries as good human facilitators would be. Achieving this will likely require further innovation, user testing, and perhaps new standards (much like privacy regulations) around recording consent. It's a challenging problem, but solving it is key to **human-centric AI**. By ensuring every user's "yes" is meaningful and every "no" is honored, we pave the way for technology that genuinely *serves* users – fostering interactions that are not only intelligent, but also empathetic, safe, and worthy of our trust.

Sources: [4] [34] [12] [14] [9] [10] [11] [16] [17] [18] [19] [20] [21] [31] [32] [33] [30] [1] [3] [2]

---

1 Constitutional AI: Harmlessness from AI Feedback \ Anthropic
https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

2 4 5 6 7 12 13 22 23 24 25 26 27 34 Consentful Systems
https://consentful.systems/

3 17 18 Toward Needs-Conscious Design: Co-Designing a Human-Centered Framework for AI-Mediated Communication
https://arxiv.org/html/2508.11149v1

8 9 10 11 14 15 16 Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms
https://imjane.net/papers/chi2021-affirmative-consent.pdf

19 How is informed consent related to emotions and empathy? An exploratory neuroethical investigation - PubMed
https://pubmed.ncbi.nlm.nih.gov/21393363/

20 Guideposts for an Institutional Framework of Consensual Dispute Processing | Office of Justice Programs
https://www.ojp.gov/ncjrs/virtual-library/abstracts/guideposts-institutional-framework-consensual-dispute-processing

21 31 32 On the Complexities of Enabling Demonstrated Consent – Center for Informed Consent Integrity
https://ge2p2global-centerforinformedconsentintegrity.org/2025/05/01/on-the-complexities-of-enabling-demonstrated-consent/

28 29 30 33 The Dating Metaverse: Why We Need to Design for Consent in Social VR - PubMed
https://pubmed.ncbi.nlm.nih.gov/37027706/