Trust, safety & content moderation: The coming storm

How to moderate conduct – and not just content – in a new peer-to-peer platform world.

Webhelp

Think Human

INTRODUCTION

Web 3.0 is already changing how people interact with the internet. With each iteration of the technology, the experience feels less like technology and more like the real world – one unbound by physical limitations.

Think of a concert experience. You might not be able to see a band you love playing on the other side of the world, but you can attend in virtual reality. The experience will feel richer and more participatory than streaming it on TV. It will recreate the feeling of attending a concert with 20,000 fellow fans and offer up-close opportunities to interact with them.

Consumers want this experience to feel natural: much like a regular concert, they'll walk to places they want to visit and buy T-shirts only available at the concert.

When they get lost or need help, they'll look around for human staff, spotting them in their yellow vests and uniforms. It's a deeply human, rich experience that will allow people more opportunities to experience concerts, sporting events and imagined worlds.

A new economic opportunity

It's not just the users who will benefit from the richness of Web 3.0. Brands and independent creators stand to profit. Decentralized technology is enabling more peer-to-peer transactions, creating opportunities for independent creators to monetize things like creating video game variations and NFTs. Web 3.0 platforms will take in a lot of money, but they also promise to democratize who can share in those economic possibilities.

Web 3.0 is a huge opportunity for brands – even ones that aren't traditionally digital – to earn money on digital assets, try new ideas, and market in a novel and highly engaging way.

What brands need to get right: Trust and safety

Consumers crave experiences that are engaging and intuitive, and they also need to feel safe from violations and fraud. But this critical need – paired with the exponential complexity the metaverse will introduce – is a perfect storm for brands who are already struggling with Web 2.0 moderation. Well-publicized and brand-harming moderation failures already show up in misinformation campaigns, racist and sexist AI models and catfishing.

People won't spend time and money on Web 3.0 experiences if they don't trust they're dealing with reputable sellers, think their privacy won't be protected, or fear harassment and other violations.

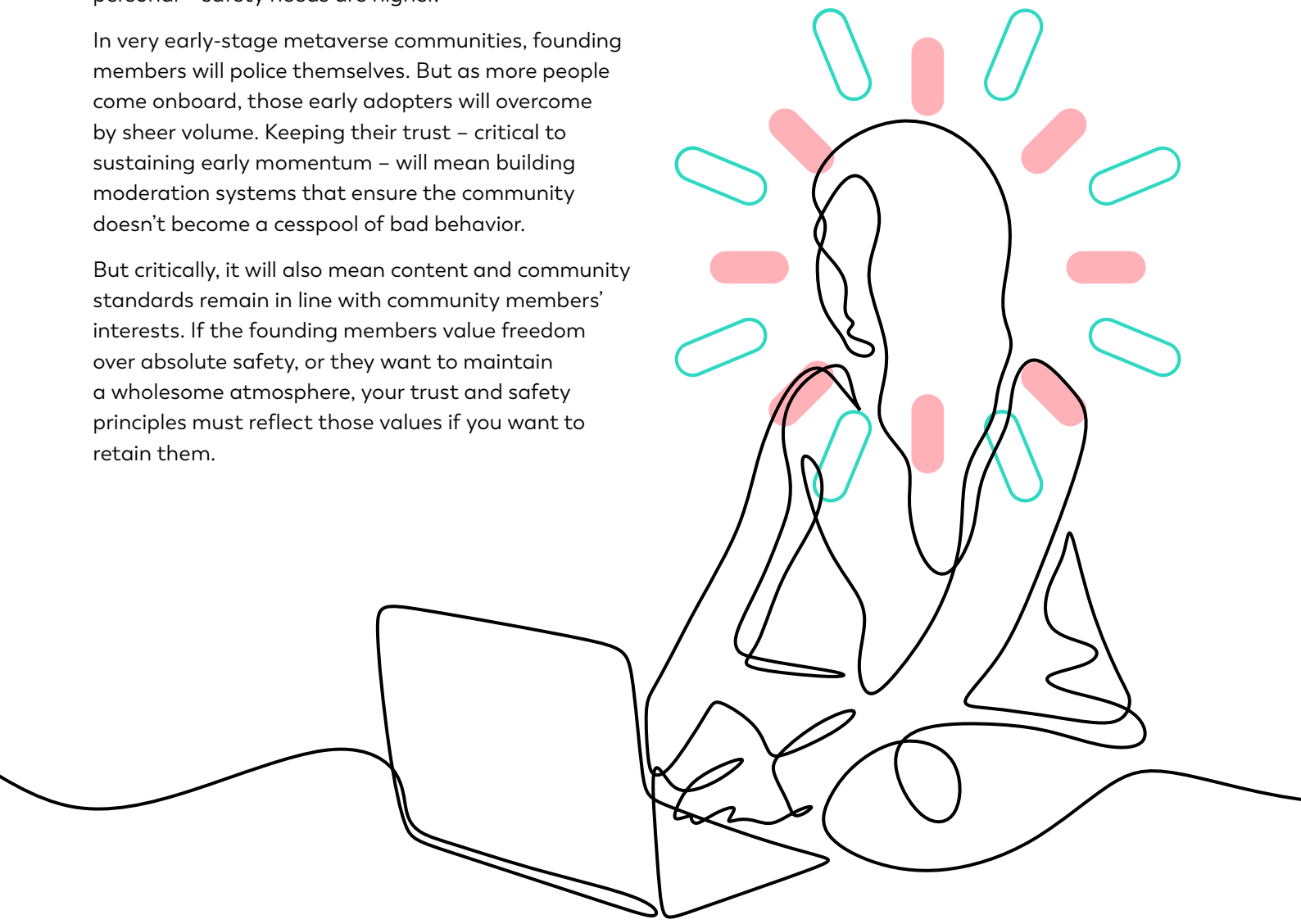
A Web 2.0 leader, AirBnB, shows the value of building trust and safety. In fact, validating the identity of property owners and renters, and providing other forms of trust and safety assurance, is the crux of their business model.
Vacationers always had the option of renting from people on Craigslist, but the vacation rental market didn't take off until trust and safety were assured.

It's a delicate balance between the needs of the core community and the new adopters who will drive profit. The brands that don't bake thoughtful trust and safety into their experiences won't reach that critical mass. And it's likely that sensational transgressions, such as virtual sexual assaults on avatars, will be widely publicized – making it harder to attract and re-engage participants.

Brands should look at trust and safety as a value driver that enables adoption, not a cost center.
And in the metaverse, with its deeper human engagement – a feeling of being up close and personal – safety needs are higher.

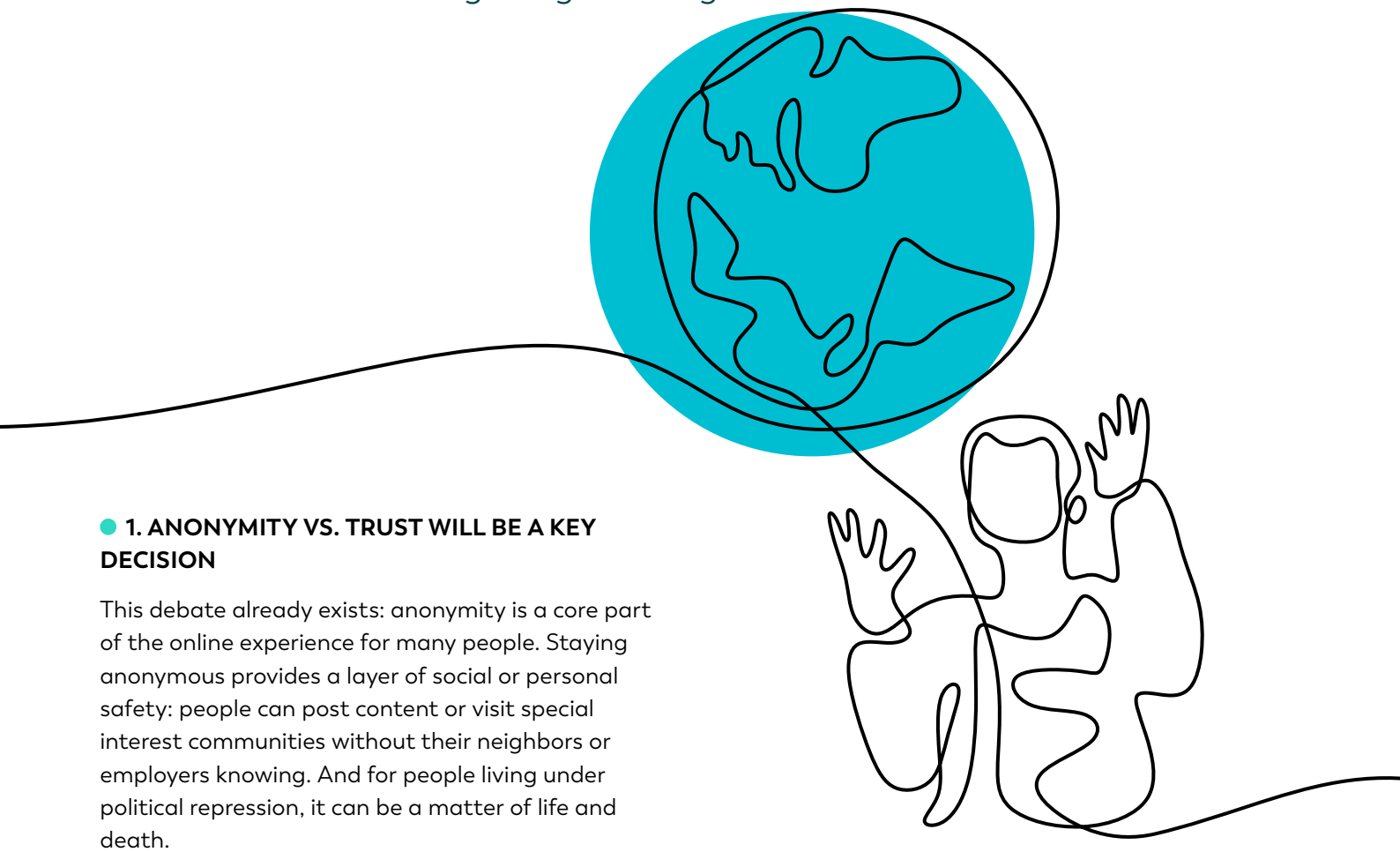
In very early-stage metaverse communities, founding members will police themselves. But as more people come onboard, those early adopters will overcome by sheer volume. Keeping their trust – critical to sustaining early momentum – will mean building moderation systems that ensure the community doesn't become a cesspool of bad behavior.

But critically, it will also mean content and community standards remain in line with community members' interests. If the founding members value freedom over absolute safety, or they want to maintain a wholesome atmosphere, your trust and safety principles must reflect those values if you want to retain them.



5 EMERGING TRUST AND SAFETY **PRINCIPLES IN THE METAVERSE**

— Every community, platform and experience has different requirements, but a few themes are beginning to emerge.



1. ANONYMITY VS. TRUST WILL BE A KEY DECISION

This debate already exists: anonymity is a core part of the online experience for many people. Staying anonymous provides a layer of social or personal safety: people can post content or visit special interest communities without their neighbors or employers knowing. And for people living under political repression, it can be a matter of life and death.

At the same time, identity is a key factor in building trust in transactions: just like in sharing apps, knowing whom you're dealing with is important for seller and buyer. Verifying identity may be important for facilitating peer-to-peer transactions. One way companies may address identity is a blockchain representation.

Here, a user's identity is represented by a non-fungible token that proves its veracity. Users can choose how much information to tie to that token to retain control of their personal safety. Proven identities drive trust in the platform and disincentivize users from actions that might disqualify them from entering other platforms.

2. SAFETY IS SIGNIFICANTLY MORE NUANCED – AND HARDER TO POLICE – IN THE METAVERSE

People feel less psychological distance in a medium that replicates the sensations of the physical world. Simulated in-person interactions offer more opportunities for violation than just trolling and hate speech. For instance, a player might make inappropriate gestures via their avatar or follow someone around.

It's also harder to use AI to police interactions and conduct than content – for one thing, the algorithms, datasets and models haven't been developed for this use case yet. And the difference between waving and an epithet might be subtle, and other variables like repeating gestures or making them in certain contexts, can change the meaning.

Cultural differences matter here, too. What seems comfortable and normal in Sweden is likely to be different in more religious cultures.

3. ESTABLISHING SAFETY WILL NECESSITATE MORE HUMAN MODERATORS

The high-volume interactions, nuance and real-time nature of metaverse interaction means there will be a much greater need for human moderators. Metaverse interactions happen in real-time, which means companies can't just get to it when they have people available. You can't turn back time and block an interaction. There's also a scale problem: moderating every interaction for 20,000 concert-goers over three hours is a lot harder than moderating the handful of tweets they might post.

As a rough estimate, metaverse events could require a content moderation ratio of something like 1 moderator for every 35 users. It's a daunting number, but brands may be able to draft community members to help with spikes in demand, possibly even paying them with free digital perks like special avatar icons. Brands also need to play for how to get double duty out of those moderators: over time, their input can help build and refine data models. Over time, that data can be used to develop AI as a first line of defense to reduce human headcount.

4. CONDUCT MODERATORS WILL NEED DIFFERENT TRAITS AND SKILLS THAN CONTENT MODERATORS

Moderators – or experience ambassadors – aren't just police, they should provide assistance: a friendly face to guide users through technical questions such as how to customize an avatar. That means moderators need to be familiar with the platform or game.

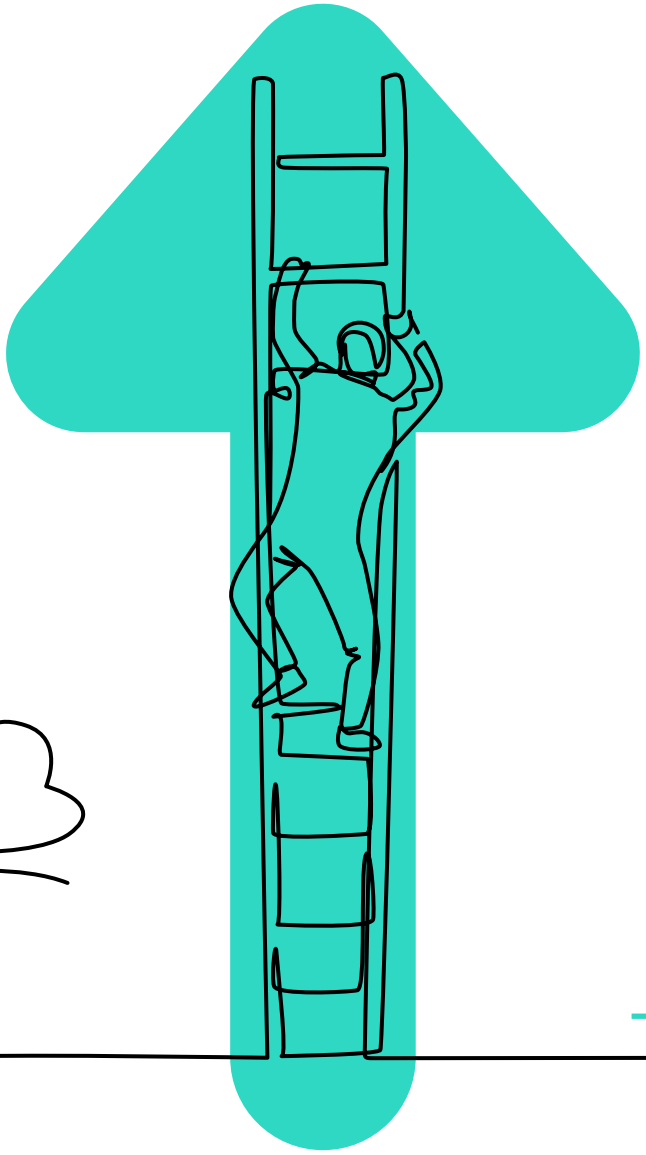
Moderating conduct is a more social job than content, so ambassadors need to be comfortable interacting. But since they won't be exposed to a high volume of disturbing content, mental resilience is less important than with content moderators.

5. FAILURES ARE INEVITABLE – LEARN HOW TO WEATHER THEM

With so many interactions happening in so many places – including universes independent creators construct – violations will occur. And users will also push back on moderation that seems inconsistent.

The physical world offers some clues on how to handle these issues. Consumers don't tend to penalize an airline if a passenger punches another one, but they will have questions if the airline doesn't handle it properly. Or to take a different example, airlines have recently gotten bad publicity for monitoring passengers' clothing choices. In both cases, having procedures that your people consistently apply can keep a minor incident from sparking a publicity fire.

Brands must build robust and transparent standards and procedures for supporting trust and safety. And those standards should be published online so users, creators, press and your employees can access them.



The stakes are high

Things are moving quickly. The pandemic pushed the world deeper into screens and heightened people's interest in interactive social experiences, and this has accelerated Web 3.0 adoption.

That signals a transformation in how people spend time and money. As virtual experiences become a more important way to experience the world, people will spend more money on them.

Richly interactive Web 3.0 platforms with peer-to-peer elements (like Roblox) are already helping companies earn real money from inexpensively produced virtual assets.

It's also helping innovators try out new ideas. When retailer Forever 21 saw their virtual hat go viral, they created a physical version, enjoying the luxury of a built-in audience and virtually no R&D cost or risk.

Marketing stands to benefit, as well. Clothing stores are already considering how they can use VR to connect more deeply with customers, such as enabling a realistically sized avatar to see how clothes fit and look—without stepping foot in a store.

But it all hinges on trust and safety. Companies that wish to claim the early-mover benefits and start creating and testing new experiences will need to create robust trust and safety programs now. The real-time, high-volume nature of conduct moderation—and the fact that children are likely to use these experiences—raises the stakes and complexity.

Weathering the early challenges will require building thoughtful trust and safety guidelines and robust conduct and content moderation teams.

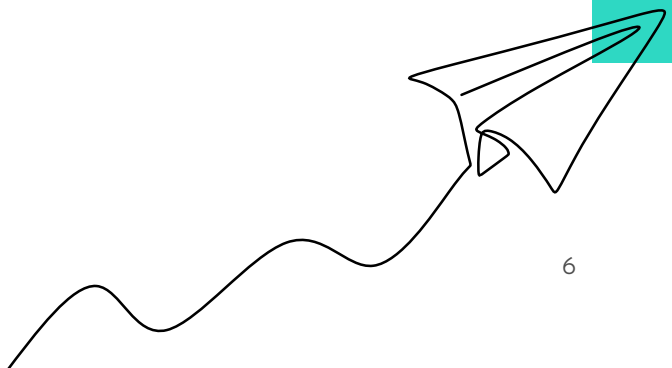
We can help you with that. Webhelp has unparalleled experience in supplying digital content services (including content moderation, digital annotation, content and community management and digital activation) to clients across a range of industries and geographic regions.

And, yes: We've done it all in the metaverse, too.

That's why we could be the perfect partners to help you ride out the incoming trust and safety storm and come out way ahead in the race for metaverse success.

So drop us a line. We're ready when you are.

GET IN TOUCH



Webhelp designs, delivers, and optimizes unforgettable human experiences for today's digital world – creating game-changing customer journeys.

From sales to service, content moderation to credit management, Webhelp is an end-to-end partner across all B2C and B2B customer journeys.

Its 100,000 passionate gamechangers across more than 55 countries thrive on making a difference for the world's most exciting brands.

Webhelp is currently owned by its management and Groupe Bruxelles Lambert (Euronext: GBLB), a leading global investment holding, as of November 2019.

Webhelp

Think Human