HARVARD BUSINESS REVIEW

Magazine Article / Product Development

Balancing Digital Safety and Innovation

How user-centric product design can enable both. *by Tomomichi Amano and Tomomi Tanaka*

From the Magazine (May–June 2025) / Reprint R2503L

Mauco Sosa

Three years after her husband died, Alice joined a dating app. The platform, driven by advanced algorithms, prides itself on being able to connect people who might never meet otherwise. Alice soon began chatting with Jim, who lived across the country. He empathized with Alice's grief and loneliness because he'd recently lost his partner too. Six months after they connected on the app, Jim was laid off. He became deeply depressed. He fell into debt. Alice tried her best to console him.

She even sent him money. But a few days after her check was cashed, Alice was devastated to discover that Jim had deactivated his account. After a police investigation, she learned that Jim wasn't real: Scammers had crafted a persona to exploit her—and she wasn't the only victim. The scammers targeted vulnerable users, exploiting the very ease and scale that made the service so appealing and innovative.

Dating platforms are hardly the only digital product with this type of vulnerability. As companies race to introduce digital features and products, they often ignore potential risks. Even seemingly benign digital products pose unforeseen dangers that can enable horrific criminal conduct. Consider Telegram, a messaging app that promotes its encryption and anonymity features. In 2019, a South Korean man named Cho Joo-bin used the platform to lure teenagers and young women into revealing personal information. He then used it to blackmail them into sharing lewd personal videos, which he sold online. In 2020, authorities sentenced Cho to 40 years in prison.

As consumers' interactions with digital products and platforms increase, more companies are recognizing the need to focus on the safety of their digital products from the earliest stages of development. Product development that considers safety as a key feature is, in and of itself, not novel. In industries like automotive and aviation, safety is prioritized during product design. But designers of consumer-facing digital products have tended to focus on novelty and speed—"move fast and break things"—creating a product development culture in which customer well-being has not been a central concern. As digital products increasingly involve personal connections, payments, and our most private information, that view needs to change—as does the way these products are developed.

We are advocating for a new model for product development. We've spent more than a decade researching safety and consumer behavior in relation to digital products, teaching online safety to more than 200 professionals, and delivering lectures on the topic in venues such as the United Nations and Harvard Business School. We contend that companies developing novel digital products must embed safeguards into their very core—starting in the earliest parts of the design process. They must establish a road map for continued improvement and open a dialogue with customers. Safety should not be an add-on but a fundamental product feature. A product's safety must apply to all users, including the most vulnerable. It may sound like we want to place an additional constraint on already highly burdened product-development teams. But the model we recommend, safety by design (SBD), can facilitate innovation rather than constrain it by providing a principled approach to product development.

A Three-Step Model

Savvy companies implement SBD using a three-step model: They align organizational priorities around safety, embed safety considerations into product development processes, and foster continuous user engagement and feedback.

Align around safety. Organization-wide support for a focus on safety is crucial. Smart companies talk about safety as a core priority and explicitly include it in their missions and values. Consider how OpenAI describes its approach to product development in its publicly available charter:

> *We are committed to doing the research required to make AGI [artificial general intelligence] safe, and to driving the broad adoption of such research across the AI community. We are concerned about late-stage AGI development becoming a competitive*

> *race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project. We will work out specifics in case-by-case agreements, but a typical triggering condition might be 'a better-than-even chance of success in the next two years.'*

Another way an organization demonstrates safety as a core value is by embedding it into its organizational structure. This could mean elevating an executive in charge of safety to the C-suite level or taking other steps to increase the visibility and power of safety on the org chart. For example, at OpenAI four safety-related teams—policy, safety, security, and red teaming (a group that simulates cyberattacks to assess security)—engage throughout the product development process. They work collaboratively to identify and address potential risks associated with AI technologies. Because of its commitment to responsible AI development, the company prioritizes engineering and machine-learning research on urgent risk issues.

Companies must also communicate their commitment to safety to customers and external stakeholders. They should point to specific tools and design choices that deliver on this commitment. When users see that their well-being is a priority, they start to trust the organization, believing that it is actively working to protect their interests even while acknowledging that no product is without risk.

Companies should communicate not just their intentions and commitments but also their results. For example, Microsoft Xbox publishes transparency reports twice a year, which show enforcement statistics related to its community standards. The reports demonstrate that Microsoft is serious when it asks users to abide by them. By sharing the company's learning trajectory and demonstrating its progress,

Microsoft helps consumers participate in the ongoing dialogue about safety risks in gaming communities.

Embed safety into product design. Just as market research provides foundational data for product developers, companies that utilize the safety-by-design model use risk assessment frameworks at the earliest stages of product conception to better understand vulnerabilities and identify opportunities to reduce or eliminate them. The design philosophy calls on product managers and engineers to address potential risks proactively instead of putting products in the market and waiting to see how users might exploit them.

Risk assessment should evolve over the life of a digital product. An SBD team must search for emerging user needs and trends, often in close communication with user-experience teams, legal teams, and even end users. This step of SBD works because it empowers those with a vision or a concern about risk to share it.

Foster continuous engagement. Communicating about safety helps build trust with users and encourages them to engage on safety issues. A company can enhance engagement by asking users and those in the community to use its product in a way that is aligned with its vision. For example, Pinterest takes a unique approach with its Creator Code, which asks users to "express yourself…but be kind." Instead of focusing solely on prohibited behaviors, the code outlines specific positive behaviors it encourages on the platform. Pinterest also highlights support for the code from some of its leading creators, thereby increasing the legitimacy and effectiveness of the initiative.

> **When a company prioritizes safety within its organization and its product development flow, users tend to notice.**

A company can also drive engagement by communicating regular progress updates as well as the cumulation of its learnings and changes over time. Doing so reinforces the idea that the company itself is willing to learn, both from its mistakes and its community, which in turn encourages consumers to deliver feedback. Consider Pinterest again, which publishes a timeline of its initiatives that showcases the company's ongoing commitment to enacting policies that protect users. Airbnb provides another example of this principle in action. The company published a narrative history of its ban on parties, detailing the evolution of its policies and the reasoning behind them. Helping users understand the context of rules may increase their willingness to comply.

Companies that follow through on commitments, share what they learn, and consistently enforce rules demonstrate that their safety initiatives are an integral part of their operations. By example they encourage users to take risk seriously and modify their behavior accordingly. Further, because their users understand that the product development process is continuously evolving, they are less likely to seek a "perfect" product; they're willing to engage with a company that demonstrates its commitment to learning and to updating its products in a user-centric way.

How Companies Use the Model

The three-step model enables companies to prioritize safety without stifling innovation. Let's look at how some companies have implemented safeguards into their products while facilitating innovation, driving sales, and improving customer relationships.

Snap Inc., the parent of social media company Snapchat, has integrated SBD principles into its platform. Recognizing the potential risks associated with digital communication, Snap has made it a priority

to protect users, especially young community members, from harmful content and interactions. For example, Snapchat introduced the Here for You feature in 2020, and it provides localized mental-health resources when users search for terms associated with crises such as anxiety, grief, or bullying. Here for You effectively embeds an important safety feature into a core offering of the product—searching for videos.

Another notable example of Snapchat's safety initiatives is its Trusted Flagger program, which enables nonprofits, government agencies, and safety partners to report content that violates community guidelines. Reports from trusted flaggers are escalated immediately, allowing the company to quickly identify and remove particularly dangerous content. In fact, flagged content is often removed within 15 minutes, preventing spread and making the platform more inclusive and welcoming, even to users who are most at risk of being targets of such content. Efforts like this have helped Snap engage cooperatively with industry experts and coalitions that had previously been frequent critics of its product.

Tinder, the world's largest online-dating platform, is another company prioritizing safety in the design of its offerings. A host of product design features, such as photo verification, a dedicated fraud team, and user education tools, help mitigate catfishing scams like those mentioned at the beginning of this article. And in May 2021 Tinder launched Are You Sure?, which uses AI to detect harmful language in messages *before* they're sent. When such language is identified, the system prompts senders to reconsider their words. The Are You Sure? feature has changed how users engage, creating a more respectful environment: Users who encountered the "Are you sure?" prompt were less likely to be reported for inappropriate messages the following month. On Plenty of Fish, Tinder's sister app, there was an 84% decrease in "Are you sure?"

prompts triggered just a few months after the feature was launched. Users had adapted their communication style to be more respectful.

Mauco Sosa

The Are You Sure? program was part of a broader, multifaceted safety initiative at Match Group, Tinder's parent company. (One of us, Tomomi, is the former director of safety by design at Match Group.) The safety-by-design philosophy has become woven into the fabric of product development at Match Group. A dedicated safety team reviews all new products and features. It ensures risk considerations are addressed from conception to launch by working closely with product, engineering, legal, and privacy teams. By embedding the SBD team within the product development process instead of with the legal, regulatory, or compliance groups, Match Group demonstrates that user well-being can be a fundamental part of product development.

Enhanced product safety often promotes inclusivity by reducing the amount of negative behavior that marginalized groups experience, an unfortunate and frequent occurrence in digital culture. For instance, TikTok uses AI algorithms and human oversight to identify harassment or bullying—especially of minors, the platform's most engaged and most vulnerable users. And YouTube's safety mode restricts content to ensure that minors can't engage with or view explicit or inappropriate content. In these instances, the principles of SBD are embedded into the products through advanced detection and mitigation technologies. Researchers from Bangabandhu Sheikh Mujib Medical University determined that machine-learning-based detection systems are up to

91% effective at identifying cyberbullying and harassment behavior, and they're much faster than human detectors.

Continuous Adaptation

When a company prioritizes safety within its organization and its product development flow, users tend to notice. We know this is true because we can track customer engagement following organizational alignment on safety. Users understand that the current version of the product is not perfect, and they trust that the company will continue to refine and improve it. Further, they are also more likely to change their behavior and work with the company to enhance the product. This isn't a new phenomenon: A 2008 study in the *Journal of Interactive Marketing* showed that safety cues such as product updates, security assurances, and transparency in product development help lower users' perception of risk and enhance their engagement with products.

This feedback loop can create a virtuous circle. Contrary to concerns that safety measures hinder innovation, we've seen that SBD can drive product development: Tinder's use of AI tools to examine messages before they are sent, for example, and Snap's integration of health resources into its core search capability are novel product designs that would likely not have emerged had safety not been a core part of product development.

For another example, consider Instagram's activity dashboard, which helps users monitor the amount of time they spend on the platform. Instagram decided to address concerns about overusage and its impact on mental health through design, despite its initial concerns that the dashboard could potentially decrease how much time users stayed on the app. It gave users the ability to set reminders to alert them when they reached their daily limit on the app. Further, it designed an environment that researchers at Stanford University found is likely to

enhance user well-being as well as improve the overall user experience, leading to higher engagement and trust.

Successful SBD effectively changes user behavior. Uber Japan developed a checklist that delivery cyclists must go through before they can receive orders, and they must complete the checklist every time they log in. It involves reviewing and confirming key safety measures, such as wearing a helmet, not using a mobile phone while riding, following traffic rules, knowing when to call the police in emergencies, and selecting the correct bike route in their navigation app. (Tomomi was the lead developer of this product feature.) The checklist works because it reaches the cyclists at an ideal time—when they are about to start their deliveries for the day. The design reflects the behavioral economics principle of nudges, which are typically costless modifications to product features that can result in significant behavioral changes. The feature was subsequently rolled out in Australia and other Uber markets.

> **Companies don't need to do a massive product-development overhaul to get started. They can incorporate safety principles in small waves.**

The significant decrease in "Are you sure?" prompts at Tinder indicates that users adapted their behavior to align more closely with the platform's intended use. Match Group's efforts to promote a safe environment affected the way consumers and stakeholders saw the product. "By conveying [its] expectation for respectful communications, and letting users pause a moment to rethink a message that might offend, Tinder is engaging its community to create a safer platform,"

Scott Berkowitz, president of RAINN, the nation's largest anti-sexual-violence organization, said in a statement posted by Tinder.

Sometimes including safety in product design can lead to new lines of business. Take Airbnb's party ban, for example. After several highly publicized instances of Airbnb users hosting large parties in residential homes, the company enacted the ban and clearly communicated its rationale to users and hosts. The company explained that the safety of its community, and of the neighborhoods surrounding rentable properties, was its priority. Many hosts (and their neighbors) welcomed the ban. And the platform saw a 44% year-over-year drop in the rate of reports of disruptive parties. Still, a number of Airbnb hosts wanted to be able to welcome weddings, parties, or other large events at their properties. So the company introduced rental categories; hosts with larger properties that can safely accommodate more than 16 people can do so, but the official party ban for smaller listings remains in place. Airbnb found a way to generate a new revenue stream while also protecting the safety and interests of hosts and their neighbors.

Getting Started

Companies don't need to do a massive product-development overhaul to get started. They can incorporate safety principles in small waves. We recommend starting with a risk assessment. Begin by posing questions that are tailored to your particular context. These questions should build on baseline questions such as, What is the risk? What are the consequences if the risk is not addressed? What is the level of the risk? What is the likelihood of the risk? Your answers will fuel conversations about what could happen when consumers of various backgrounds and preferences use your product.

While imagining potential risks may be daunting, involving product managers in the assessment will help advance this conversation.

Product managers must commit to integrating safety into the design. They should also answer key questions, such as, How will the risk be monitored? When will the risk be next assessed? Having an iterative back-and-forth with the product manager is consistent with the spirit of SBD, and it helps expand the scope of risks that may not have been apparent initially or were hidden in the fog of the product development process. Other stakeholders, including corporate legal and privacy teams, are also helpful resources to flesh out additional risks and challenges.

The risk assessment should reflect the broad use cases for the product. It is impossible to imagine every possible way in which consumers might use (or misuse) a product, but it is helpful to adopt an expansive view from the start.

The presence of risk doesn't mean that a new product or feature shouldn't be introduced. Rather, it is a starting point for a dialogue among the safety advocate, product managers, and other stakeholders. One concrete step to take is to identify the relative priority levels among the identified risks, given the timeline and resources. A product team can then decide whether to take action immediately to mitigate a risk or to accept a risk for the time being.

Another important step is to set up systems to continuously monitor and address online threats and vulnerabilities. Risk assessment should be seen not as a one-time event but as an ongoing commitment to product safety. And so these systems must continually be updated over time.

Some organizations have constructed templates for risk types by accumulating knowledge from past assessments. They then taxonomize the risks inherent in new products to streamline the assessment

process. For instance, an online product that involves user-generated content might have risk categories relating to "user-generated content," "misconduct involving minors," "offline harm," and "algorithmic bias." The goal here is not to attempt to cover all possible risks (which is impossible to do) but rather to have a grounded approach to risk assessment.

• • •

The safety-by-design model doesn't require a complete overhaul of a company's processes. Even minor adjustments derived from a risk assessment can gradually spill over to more-considerable organizational changes. That might manifest as a change in the company's overall product-development workflow or as a new revenue stream, as we saw with Airbnb.

SBD does require continuous learning, adapting, and improvement. Companies that embed SBD into product development and organizational culture will build trust with their users. They will also improve the user experience, which will lead to higher engagement—both of which contribute to long-term success.

A version of this article appeared in the May–June 2025 issue of Harvard Business Review.

TA

Tomomichi Amano is an assistant professor of business administration at Harvard Business School.

TT

Tomomi Tanaka is the founder of Safety by Design Lab and an adjunct professor at Kindai University in Osaka.