PARTNERSHIP ON AI

RESPONSIBLE PRACTICES FOR SYNTHETIC MEDIA CASE STUDY

# How Bumble is preventing malicious AI-generated dating profiles

bumble

This is Bumble's Case Submission as a Supporter of PAI's Synthetic Media Framework.
Learn more about the Framework

# 1 Organizational Background

A contextual introduction to the case study.

### BUMBLE'S RESPONSE

Bumble Inc. is the parent company of Bumble, Bumble for Friends, Badoo, Fruitz, and Official. Through Kind Connections™, Bumble Inc.'s platforms enable people to build healthy and equitable relationships.

Founded by Whitney Wolfe Herd in 2014, Bumble was the first dating app to focus on women's experiences. Bumble connects people across dating (Bumble Date), friendship (Bumble For Friends), and professional networking (Bumble Bizz). Badoo, which was founded in 2006, is one of the pioneers of web and mobile dating products. Fruitz, founded in 2017, encourages transparent communication about dating intentions through playful fruit metaphors. Official, founded in 2020, is an app for couples that promotes open and honest communication between partners.

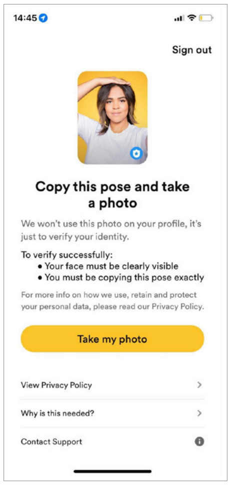
Bumble Inc. takes the safety of all its members very seriously and prides itself on being a space where kind connections can be made in a safe, inclusive, and respectful way. In particular, we make it clear in our Community Guidelines that we prioritize fostering a community built on genuine connections and that inauthentic profiles are prohibited. Bumble celebrates authenticity, and we expect all our members to represent themselves accurately on their profile.

The Trust & Safety Collective at Bumble Inc. is a cross-company team that consists of members from the Safety Policy, Product, Operations, Engineering, and Data Science teams. The Collective's agenda is to ensure that members of the Bumble Inc. community feel safe and confident while using our products.

As part of our commitment to authenticity, we launched Photo Verification on Bumble in 2016. This feature helps confirm that the photos on a member's profile match the person using the account. In addition to Photo Verification, Bumble Inc. also uses a combination of automated technology and skilled human moderators to proactively detect and flag potentially fake profiles within the app. This team of moderators may block members or request Photo Verification when they investigate whether a profile is suspicious. In order to get their profile photos verified, members must take a selfie mimicking a specific pose, pulled randomly from 100 example poses, provided by Bumble. Once they submit their picture, both automation tools and human moderators will review the image and the member will be notified within minutes if their profile is confirmed for verification. If they're verified, a blue verification badge will be added to their profile. If they don't get verified the first time, but nothing malicious is suspected, the member will be able to try the process again until they're verified.

Our mission has always been to enable a safe experience for connecting with people online. Our Photo Verification technology was initially rolled out to keep our members safe, but requesting verification from a potential match can add an extra layer of confidence to every interaction. This means that members can connect with each other knowing that the profile photos match the person using the account.

Our Photo Verification technology relies on comparing the member's profile pictures with the selfie that the member is prompted to take as part of the verification process. It would be extremely hard for a member to know which pose they'll be given to depict in the selfie before the verification process starts. The selfie, therefore, serves as evidence that there's a real person using the device to take the photo. This method worked well for years. It allowed Bumble Inc. to continue evolving as a secure space for members to connect, while minimizing friction by not requiring members to share personal documents or other forms of identity verification.



**Copy this pose and take a photo**

We won't use this photo on your profile, it's just to verify your identity.

To verify successfully:
- Your face must be clearly visible
- You must be copying this pose exactly

For more info on how we use, retain and protect your personal data, please read our Privacy Policy.

**IMAGE 1.** A screenshot depicting the prompt for Photo Verification on Bumble.

Huge advancements in synthetic media have resulted in it being easier than ever for folks to create realistic and fully customizable images of non-existent people. Since these developments, we've started to see an increase in potentially harmful attempts to create profiles and garner photo verification with synthetic images. These actions are considered potentially harmful based on the impact that they'd have on our platforms and for our members. If these attempts were successful, not only could they potentially cause harm to our members, but they'd also hurt the trust that our members have in the Photo Verification feature, and in verified profiles.

Due to the increased quality of these synthetic media generation technologies, bad actors are able to create photos that are almost (if not completely) indistinguishable from legitimate ones to both automated and human moderators. Photo Verification, while leveraging human intervention throughout the process, is powered by pixel-level machine learning (ML) models. This combined feat allows for smooth operation at scale. However, the more we see synthetic images start to fool humans due to their quality and realism, the less reliable ML systems (even the specially trained ones) might be in assessing facial similarity.

**IMAGE 2.** Pictures depicting a synthetically generated person (R + L), mimicking the Photo Verification process (R).

If these falsified images were verified, Bumble Inc. could be seen as a **Passive Distributor** of synthetic media—both in cases when someone is trying to bypass our verification systems with a potential intent of harm, and when someone is trying to depict themselves (or non-existent people) in counterfeit situations. Having the bad actor's profile verified could increase their popularity on the app, allowing them to potentially gain more matches and maximize the reach of their potentially harmful intentions.

From a safety and product policy standpoint, the role of Photo Verification on our platforms—and how intrinsic the feature is to our mission—has been clear. This is reflected in the sanctions we have in place for attempts to bypass it (e.g., we might block these profiles for being inauthentic). Aside from potentially harmful malicious uses of synthetic media, there's another complex policy conversation to be had around the use of generative AI in a non-malicious capacity on our platforms. The policy work here revolves around assessing the appropriate balance between authenticity and creative freedom and taking into account the opportunities for non-malicious self-expression that synthetic media can offer.

## 2 Challenge

Elaborate on the challenge being addressed in the case study, i.e. the issue to which your organization is applying the Framework.

### BUMBLE'S RESPONSE

At Bumble Inc., we first noted an increase in profiles being created with synthetically generated images in the second half of 2022. This was due to huge advancements in generative AI, which made it easier than ever for people to access this kind of technology and create photorealistic, fully-customizable images of nonexistent people.

We began to see an increase in fake profiles attempting Photo Verification using both

images of popular figures, influencers, and other celebrities and AI-generated images. These use cases would fall under Appendix B of Partnership on AI's (PAI) Framework — "Impersonating an individual to gain unauthorized information or privileges."

These scenarios were identified by different subsystems and components of our Trust & Safety Collective. For example, in our dashboards, we saw an increase in verified profiles that were later reported and blocked for being inauthentic. In our automated Quality Assurance (QA) and manual moderation queues, we started noticing more cases of potentially harmful profiles seeking verification (sometimes successfully) with pictures that were later confirmed to be inauthentic and possibly generated by AI.

These patterns could pose potential harm to our members. Users of Bumble Inc.'s products have increased trust in profiles that are verified, and may rely on our certification of authenticity. Bumble Inc.'s apps aim to help our members foster kind connections, meaning that those who create malicious profiles using synthetic images go against our mission. They also act against our community guidelines by misrepresenting themselves on the platforms. We strive to protect our members from potential harms such as catfishing and romance scams.

As potential **Passive Distributors** of synthetic media, we had two challenges:

- Finding reliable ways to detect occurrences of synthetic media (or highly-tampered content).
- Developing safety policy approaches to deal with non-malicious, synthetically-generated content that our members upload to express themselves in a kind and healthy way.

The second point was particularly complex because of the technological advancements in synthetic media generation, together with its mainstream adoption and ease of access. These developments made the task of reliably detecting when someone was using synthetic imagery almost impossible, even to human eyes.

The most promising initiatives in this area revolve around secure media provenance approaches, such as the C2PA standard. These standards depend on secure and encrypted manifests that meticulously track an image, video, or audio file from its inception through every subsequent edit. This process is effective when using hardware and software that support the standards, resulting in a fully observable and tamper-proof record of the entire creation and editing history of a piece of media. This solution would solve the detection issues outlined above and establish trust in the image at every step—all the way from creation to when it's uploaded on a platform. However, this approach would require industry-wide support in order to reliably use it, as well as an invaluable and forward-thinking proof of concept. This is especially true for platforms like ours which are neither **Creators** nor **Active Distributors**, per the Framework, of this type of media. In our case, the absence of a C2PA manifest alone isn't a reliable indicator of potentially harmful synthetic media or the presence of synthetically generated content.

More recently, Google DeepMind introduced SynthID, a slightly different approach to identifying synthetically generated imagery that's based on pixel-level watermarking. This technology embeds a digital watermark directly into the pixels of the image, making it detectable through systematic identification but imperceptible to the human eye. This is another future-proof and promising method but, unfortunately, it still would not solve the issue at hand as the vast majority of synthetic media is not created using generative AI, but more traditional image editing software.

However, even if there was a reliable method to detect synthetic media, we would need to adopt a reactive policy to regulate and monitor its adoption and impact on our platform. We'd also need a disclosure mechanism, such as labels (once carefully designed and tested), in order to iterate and experiment.

The policy effort would be particularly nuanced within our business and industry, which revolves around dating and connection apps. This is because it involves navigating the intricacies of personal expression and creativity. It would also mean we'd be early adopters of safety policies in this area, with limited past experiences or literature to draw from.

# 3 Objective

Describe what your organization is attempting to accomplish by addressing this challenge and/or furthering the opportunities.

### BUMBLE'S RESPONSE

Our goals for implementing the PAI Framework's principles into our products were primarily to:

1. Further mitigate the risk of potential harm resulting from members relying on a user's verification status (or lack thereof) and the potentially harmful use of synthetic media. This would be achieved by employing direct synthetic media detection or leveraging other enhanced behavioral technologies to spot bad actors. Ideally, these technologies should identify bad actors before or immediately after their attempts to circumvent our systems using synthetic media (see next section).
2. Conduct an additional investigation to identify more reliable ways to detect synthetic media for either potentially harmful or non-malicious purposes. In light of recent technological developments in the area, this investigation would allow us to build foundational knowledge that would put us in a strong position for years to come.
3. Investigate, design, and roll out synthetic media policies to aid our efforts in informing our members that some people are uploading images that have been heavily filtered, edited, or tampered with. This is with the caveat that these images may not violate our community guidelines or have been uploaded without explicit malicious intent.

# 4 Framework Scope and Application

Identify which Framework principle was used to help address the challenge/opportunity, how it was chosen and implemented, and describe how it was applied.

### BUMBLE'S RESPONSE

When considering the above, Bumble Inc.'s products can be categorized as possible **Passive Distributors** of synthetic media. Therefore, our role is to:

- Find additional, reliable ways to detect the occurrence of both potentially harmful and non-malicious synthetic media.
- Develop mechanisms to notify our members when non-malicious synthetic media is detected—and potentially implement disclosure approaches as well as downranking measures for profiles.

Bad actors may typically use synthetic media to bypass our verification systems and create fake profiles on our platforms. To try and preempt this behavior, we looked at the possibility of spotting and potentially downranking profiles that showcase other suspicious patterns on our apps.

From a safety policy perspective, we're primarily concerned with the content of the images, and not how they are generated. For example, if someone uploaded a picture of a child on their own, this would infringe on our Community Guidelines, no matter whether it was generated with AI or if it was a photo of a real child. Either way, this would be a clear violation of our platforms' rules and would trigger sanctions. However, if someone uploaded a clear picture of an adult smiling, although possibly generated through AI, it doesn't directly infringe our guidelines, so it would be much harder to assess and deal with.

Taking a step back, our goal as a Trust & Safety Collective is to reduce the number of fake profiles on our platform. We aren't constrained by the technology we use to do it, nor limited from a policy perspective by the use of synthetic media. On the other hand, the PAI Framework prescribes disclosure mechanisms to ensure members are informed about the usage of synthetic media. Any kind of policy work or effort relating to this challenge is tightly connected with the ability to reliably detect synthetic media or manipulated content. Detecting if a picture has been manipulated or generated by AI is becoming a complex technical challenge, making the design (and enforcement) of any policy work in the area harder.

Historically, at Bumble Inc., we've always made policy decisions on the nature of uploaded content (e.g., abusive vs. non-abusive) rather than the provenance (e.g., authentic vs. synthetic).

This approach proved itself to be robust to media provenance from a member safety standpoint, but we're now reassessing it in light of synthetic media and fair disclosure mechanisms. This will allow us to build our knowledge base and carry out effective research that can inform our policy decisions and, ultimately, help our members continue to find kind, authentic connections.

# 5 Obstacles

Elaborate on any internal or external obstacles intrinsic to the Framework that were overcome.

**BUMBLE'S RESPONSE**

When having to deal with the rise of potentially harmful profiles using synthetic media attempting to bypass Photo Verification, we primarily consider how we can:

1. Continue to reliably detect bad actors at the Photo Verification stage who are trying to impersonate non-existent people.
2. Develop a Bumble Inc. safety policy approach to regulate the expressive use of synthetic media by legitimate members.

In both cases, we'd need to rely on detection technology to tell us if a specific piece of media is synthetically generated or not and, possibly, to what extent. A filtered image—which is quite common on our platform—is a different kind of content from a highly manipulated, retouched, or AI-generated selfie. Detecting them in the first place involves different complexities. This challenge led us to a wider conversation in the Trust & Safety Collective, particularly in the Safety Policy team, to develop a comprehensive or acceptable use policy for synthetic media. This would enable us to determine what would be an acceptable and authentic way for a member to use synthetic media to express themselves on our platforms.

Addressing the first topic in its entirety was more complex than we expected. However, we did successfully address the original threat we were working on: potentially harmful attempts to bypass Photo Verification. It's getting more and more complex to detect synthetic media, especially content generated by generative AI technologies (e.g., StableDiffusion XL and Midjourney). Current industry efforts to address these (e.g. C2PA, CAI, and SynthID) are extremely solid, well-intentioned, and future-proof for the next decade.

However, because of their early-stage adoption, they cannot be fully relied upon to be a high-recall detection technique that is directly relevant for our use case, especially with respect to potentially harmful attempts. For example, the "chain of custody" approach—with cryptographic provenance used by the C2PA—only works if all the organizations involved in dealing with a piece of synthetic media are able to work with it and keep it up to date. This is a compelling vision, but not practical without a joint effort from various players in the industry. This limitation makes any policy effort regarding disclosure mechanisms theoretical. This does not mean they're not worth having, but these mechanisms depend on the ability to detect synthetic media in the first place. In particular, it would be beneficial (if not fundamental) if all the biggest players in the AI image generation space would implement C2PA provenance by default on all media generated through their platforms. Such an achievement would allow us to be more prescriptive over the presence of a provenance check, given that its absence would be more correlated to actors with potentially harmful intentions trying to manipulate it after creation.

Assessing real-life use cases of attempts to bypass our verification process through synthetic media adoption allowed us to take time to reason over our internal definitions of synthetic media and their repercussions. They prompted us to better define our internal concept of authenticity and the methods we use to enforce and disclaim around it. Historically, we relied on the technical assurance in Photo Verification that, through gesture verification, there was a real person in front of the camera. This was a successful deployment and allowed us to minimize the effort required from our members to get verified, while ensuring a safe and trusted community.

We've also implemented ID verification in Japan and are planning to explore ID verification and other enhanced verification methods in different markets. These methods could also help us reduce the impact of synthetic media attempts to bypass photo verification to some extent.

# 6 Benefits

Identify the opportunities created for your organization by utilizing the Framework to address the challenge.

### BUMBLE'S RESPONSE

Having to address the issues above with the Framework in mind allowed us to ask ourselves the right questions and set ourselves up for success on reliable trajectories.

In light of the above, we decided to start from a detection standpoint. For Bumble Inc., the surge in synthetic media was not widespread nor on a product level. Rather, it appeared to be directed at a specific system (Photo Verification), likely by the same potentially harmful profiles that have historically used other techniques. This is why in this context we started to use behavioral science and ML. We began researching, designing, and deploying advanced ML technologies with the aim to further decrease the number of active members seeing a spam profile, and the number of spam profiles being liked. This new solution succeeded in the empirically verified assumption that no matter the technological breakthrough bad actors employ, they leave behavioral traces, such as IP addresses, devices, and patterns such as swiping speed/frequency, incoming reporting activity, and number of messages sent. From this, we can continue to learn and improve our detection technologies.

We also started to develop internal subsystems to deal with synthetic media directly. We had very good results by using hashing mechanisms in detecting more "classic" photo editing and retouching attempts. These interfaces are particularly useful—even fundamental—to our human moderators. It helps them block potentially harmful attempts when our automatic systems might not be able to detect on their own such convincing images.

As addressed above, these systems are likely to address only certain use cases. They're only able to deal with pictures where just the face is changed through either media manipulation software or AI-powered face swap technologies. However, they're an invaluable starting point to address the issue. The knowledge we collected and developed internally on synthetic media has been another invaluable benefit from this exercise. It allowed us to position ourselves at the forefront of the conversation with respect to C2PA and CAI, and pilot other commercial providers offering these technologies.

Discussing these challenges internally and externally led us to start introducing some related concepts in our policies, specifically in the newly updated (Q3 2023) Community Guidelines:

- "Profile Photos. We want your profile to celebrate your authentic self! That's why we require at least one of your profile photos to depict only you and to clearly show your full face.
- We do not permit:
  - Profile photos that are heavily distorted or contain exaggerated or unnatural digital effects to the point where it cannot be clearly determined that you're the person in the photos."

Alongside technological developments, these statements are another step to continue addressing the challenges arising out of synthetic media in general. They also allow us to be open to future iterations to address the topic more broadly. More precisely, in our policy for inauthentic profiles, we take a strong stance on artificially generated photos as a way for members to express their authentic selves:

- "We don't allow impersonation or misrepresentation on our platforms. We consider these, and similar behaviors, as inauthentic behavior. This may include, but isn't limited to:
  - Catfishing or Impersonation (i.e., creating an online persona that isn't you)
  - Using someone else's photos, artificially generated photos, or enhanced photos to deceive others."

# 7 Conclusion/Key Takeaways

A description of how implementing the Framework ended for your organization, including any lessons learned.

### BUMBLE'S RESPONSE

We started to witness an anecdotal increase in the quality of the harmful attempts to bypass our Photo Verification system almost in parallel to when we started to discuss and implement the PAI Framework internally. This has allowed us to progress in three different directions: exploring the industry landscape, enhancing behavioral ML, and refining authenticity definitions.

In order for us to adapt to these discoveries we:

- Conducted research, accumulated knowledge, and contributed to industry-wide initiatives aimed at detecting synthetic media on a large scale. This has enabled our company and products to be prepared for this technology and its associated implications.
- Designed and deployed improved behavioral technologies to detect malicious profiles employing these techniques. This allowed us to understand that behavioral ML can still play a big role in detecting potentially harmful usages of technologies and products, even if the tools they use evolve. For example, bad actors always tend to leave traces behind. World-class behavioral technologies through ML can still have impressive results, even if direct methodologies to detect their artifacts aren't yet fully rolled out.
- Discussed the potential of redefining our internal concept of authenticity and the methodologies we employ to detect and enforce it. We viewed this from both a product standpoint (e.g., a blue tick or ID verification) and our policies. We're actively working on improved guidelines to allow members to express themselves with synthetic media, as long as it's not at the cost of kindness and authenticity.

The PAI Framework is driving relevant and future-looking principles. For these principles to be effectively implemented and enforced, it's essential for all players in this space, especially **Creators** and **Active Distributors** (publishers), to make a concerted effort.

As **Passive Distributors**, we rely on a secure chain of custody involving key players in the creation, editing, and distribution of media to systematically verify their presence and inform policy decisions. This would ensure reliable and programmable checks for media authenticity.

There's plenty of inspiring work going on in the industry, as well as in the adoption of these tracing and provenance mechanisms. However, the absence of an indirect disclosure mechanism is not the sole proof that a piece of media is synthetic, malicious, or untrustworthy. This is especially problematic for the part of the funnel we sit in, making it more complex to write comprehensive policies and disclosure mechanisms that aren't just theoretical exercises.