

AI FORENSICS

Dr Paul Bouchaud

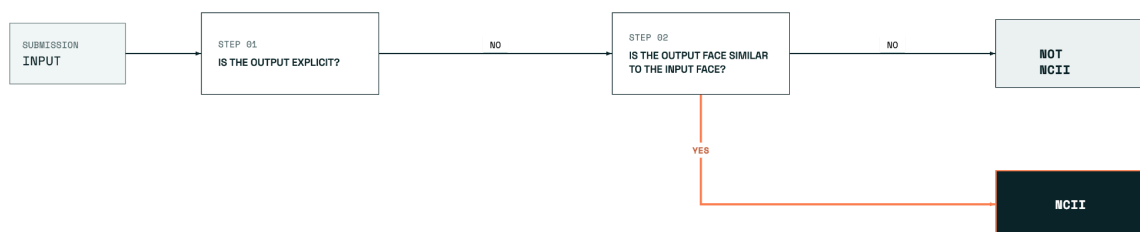
Marc Faddoul

Technical Note

Automated Detection of NCII Capability

On 28 July 2026, we published an analysis documenting the prevalence of Hugging Face Spaces (AI demos hosted on the platform) capable of generating non-consensual intimate imagery (NCII). We further documented that a large majority (73%) of the queries submitted to the image-editing Spaces we deployed as part of a live experiment were sexual in nature, overwhelmingly targeting women, and in up to 4.5% of cases, children.

Because AI-generated sexually explicit content can be produced for consensual reasons, NCII detection cannot rely on merely flagging explicit output, it requires verifying that the identity of the person, originally appearing in the input image, is preserved in the output. An output image depicting a person in a sexualized form is not NCII if that person is not a real, identifiable individual (famous or otherwise recognizable e. g. from the input image).



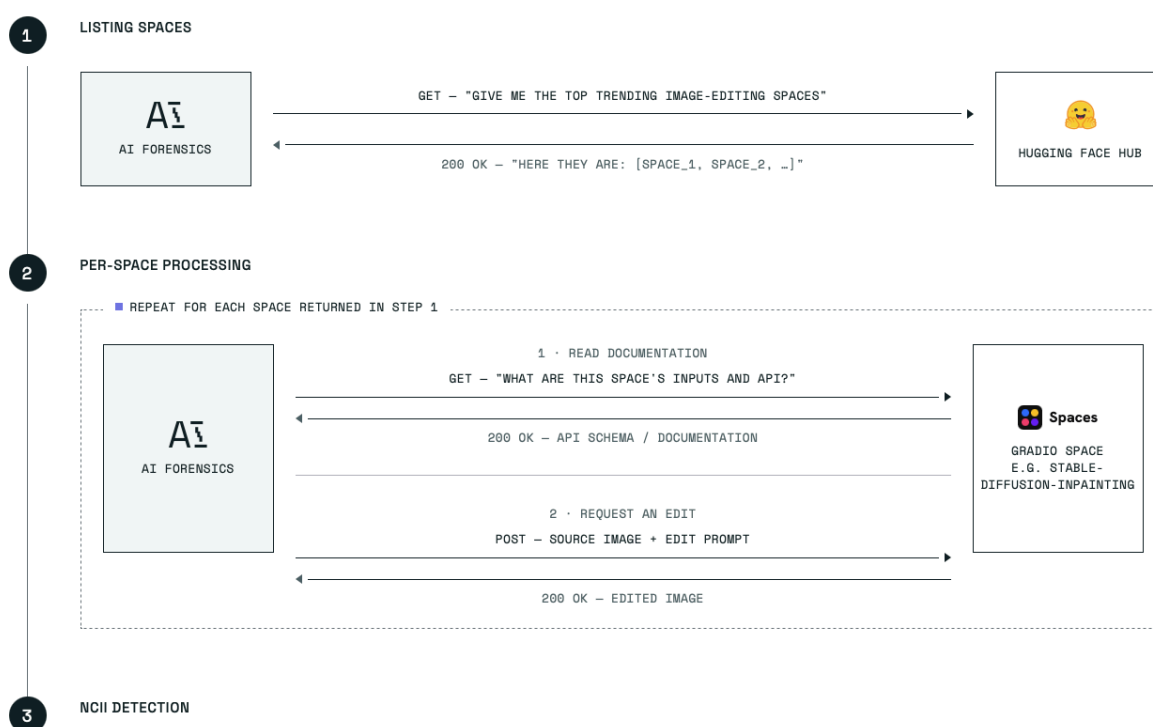
Fortunately, the tools to perform this detection are readily available: face similarity can be assessed with open-source models such as ArcFace, or through the DeepFace library, and sexual content can be detected either through body-part detection (e.g., NudeNet) or more holistically with a CLIP-based classifier.

This approach is not foolproof. It relies on an arbitrary similarity threshold between input and output faces, as well as on a definition of “sexualized content” that can differ from standard nudity detection — in our experiment, for instance, users asked to add semen to images or to place women in sexualized poses while keeping them dressed. Still, this approach offers a reasonable first filter, one that can be iterated on and refined as evasion techniques evolve.

One main limitation is that it cannot capture celebrity face-swaps, where no input image is supplied; a celebrity-recognition module could be added to cover this case. We note, however, that an overwhelming share (96%) of the NCII requests submitted to the Space we operated involved non-public individuals.

Following our study, some have argued that moderating these Spaces is not feasible because they consist of arbitrary user-generated code. Yet, according to Hugging Face metadata, 83% of the 55 most "trending" image-editing Spaces are powered by Gradio, an open-source framework that lets developers quickly build AI demos. Gradio exposes a built-in API that allows "almost any Gradio app [to be used] programmatically"¹. A large majority of Spaces — though, we concede, not all — can therefore be queried automatically.

We built and deployed an automated system that identifies image-editing Spaces on Hugging Face capable of undressing arbitrary depicted individuals, surfacing multiple such Spaces as of August 2026.



¹ <https://gradio.app/guides/view-api-page>