The UX of AI: Using Google Clips to understand how a human-centered design process elevates artificial intelligence

Josh Lovejoy
Google
lovejoy@google.com

Abstract

Google Clips is an intelligent camera designed to capture candid moments of familiar people and pets. It uses completely on-device machine intelligence to learn to only focus on the people you spend time with, as well as to understand what makes for a beautiful and memorable photograph. Using Google Clips as a case study, we'll walk through the core takeaways after three years of building the on-device models, industrial design, and user interface—including what it means in practice to take a human-centered approach to designing an AI-powered product.

Google Clips is a small form-factor camera that operates entirely offline using on-device AI. It can be stood up, held, or clipped onto things to capture candid photos of familiar people and pets.

Human-centered machine learning

As was the case with the mobile revolution, and the web before that, machine learning will cause us to rethink, restructure, and reconsider what's possible in virtually every experience we build. In the Google User Experience (UX) community, we've started an effort called "human-centered machine learning" to help focus and guide that conversation. Using this lens, we look across products to see how machine learning (ML) can stay grounded in human needs while solving for them—in ways that are uniquely possible through ML. Our team at Google works across the company to bring UXers up to speed on core ML concepts, understand how to best integrate ML into the UX utility belt, and ensure we're building ML and AI in inclusive ways.

Note that in this article, I will refer to ML as the process of training models and AI as the system architecture.

Just getting more UXers assigned to projects that use ML won't be enough. It'll be essential that they understand certain core ML concepts, unpack preconceptions about AI and its capabilities, and align around best-practices for building and maintaining trust. Every stage in the ML lifecycle is ripe for innovation, from determining which models will be useful to build, to data collection, to annotation, to novel forms of prototyping and testing.

We developed the following "truths" as anchors for why it's so important to take a human-centered approach to building products and systems powered by ML:

- Machine learning won't figure out what problems to solve. If you aren't aligned with a human need, you're just going to build a very powerful system to address a very small—or perhaps nonexistent—problem.
- If the goals of an AI system are opaque, and the user's understanding of their role in calibrating that system are unclear, they will develop a mental model that suits their folk theories about AI, and their trust will be affected.
- In order to thrive, machine learning must become multi-disciplinary. It's as much–if not more so—a social systems challenge as it's a technical one. Machine learning is the science of making predictions based on patterns and relationships that've been automatically discovered in data. The job of an ML model is to figure out just how *wrong* it can be about the importance of those patterns in order to be as right as possible as often as possible. But it doesn't perform this task alone. Every facet of ML is fueled and mediated by human judgement; from the idea to develop a model in the first place, to the sources of

data chosen to train from, to the sample data itself and the methods and labels used to describe it, all the way to the success criteria for the aforementioned wrongness and rightness. Suffice to say, the UX axiom "you are not the user" is more important than ever.

Three ways human-centered design elevates AI

Addressing a real human need

This year, people will take about a trillion photos[1], and for many of us, that means a digital photo gallery filled with images that we won't actually look at. This is especially true with new parents, whose day-to-day experience is full of firsts. During moments that can feel precious and fleeting, users are drawn to their smartphone cameras in hopes of capturing and preserving memories for their future selves. As a result, they often end up viewing the world through a tiny screen instead of interacting using all their senses.

What if we could build a product that helped us be more in-the-moment with the people we care about? What if we could actually be in the photos, instead of always behind the camera? What if we could go back in time and take the photographs we *would* have taken, without having had to stop, take out a phone, swipe open the camera, compose the shot, and disrupt the moment? And, what if we could have a photographer by our side to capture more of those authentic and genuine moments of life, such as my child's *real* smile? Those moments which often feel impossible to capture even if one is always behind the camera. That's what we set out to build.

Clips allows you to select the perfect frame and save it as a still. In this instance, I clipped the camera onto a basketball hoop to capture the moment just before my son made a basket (middle).

Guiding the intelligence

When we started the process, the most pressing question was: if people take tons of photos but don't actually want to go back and curate them, how will we label ground truth? This is where the foundational "HCML exercise" was born: Describe the way a theoretical human "expert" might perform the task today. The theory was twofold: First, if a human can't perform the task, then neither can an AI; second, by diving deep into the methods of an expert, we can find signal-to-guide data collection, labeling, and component model architecture.

The closest approximation I could think of was a wedding photographer, so I set out to find and hire contractors using a sufficiently ambiguous job posting. The interviews discussions were wide-ranging, but primarily centered on process. I wanted to find people who were particularly adept at deconstructing the many tiny decision-forks they employed in their craft We ended up discovering—through trial and error and a healthy dose of luck—a treasure trove of expertise in the form of a documentary filmmaker, a photojournalist, and a fine arts photographer. Together, we began gathering footage from people on the team and trying to answer the question, "What makes a memorable moment?"

It's important for us to recognize the amount of nuance, aesthetic instincts, and personal history that we often take for granted when evaluating the quality of our photos and videos. For example, I crack up every time I watch my younger son exploring the subtleties of a twisty straw (far left) or trying to juke my kisses (middle). And I well up with pride when I watch my older son on his bike at the park (far right), because I remember that day as a turning point in his self-confidence to ride on his own.

Building trust

The starting point for our work was an assumption that we could 'show' the model the stuff we thought was beautiful and interesting, and it would just *learn* how to find more. We had romanticized conversations about depth of field, rule of thirds, dramatic lighting, match cuts, and storytelling. But what I learned was that we should never underestimate the profound human capability to wield common sense; to quickly evaluate and prune the characteristics that are lacking in practical value.

These early experiments exposed crucial technical and methodological gaps that helped us reassess our assumptions about what the product could realize, as well as take stock in the unprecedented nature of the work. The reality of hand-held or body-worn video is that most of it is shaky, boring, poorly framed, or all of the above. So we shifted our paradigm from expecting ML to discover the most salient patterns—early models were actually quite fixated on things like hands close to the camera and abstract geo-

[1] InfoTrends Worldwide Consumer Photos Captured and Stored, 2013 – 2017

metric shapes—to understanding that it can only learn effectively under quite reductionist framings. Basically, we were trying to teach English to a two-year-old by reading Shakespeare instead of *Go, Dog. Go!*. This was where the myth of the AI 'monolith' crashed hardest for me; the idea that there's some singular 'intelligence' that understands all things and can generalize and transfer knowledge from context to context. We needed to reset our expectations and approach the task with far more pedagogy.

Back to basics

Consistency is the name of the game when trying to teach anything. It's why we wait as long as possible to unleash the madness of O-U-G-H (e.g. tough, through, thorough) on children when teaching them how to read and speak English. Spelling and pronouncing words like cat, bat, and sat, with their predictable "at" sounds, is so much more consistent!

With consistency comes confidence. Think about how quick—and eager—most students are to point out incongruity when a teacher provides two examples that don't seem to line up. Algorithms provide no such feedback. As far as an algorithm is concerned, everything they're shown is of equal value unless directed otherwise. For Clips, that meant we not only needed consistency between examples, but also within each example. Every individual frame needed to be representative of the specific prediction we're trying to teach it to make. And often that can come in the form of teaching it what to ignore.

Capture

We needed to train models on what bad looked like: hands in front of the camera, quick and shaky movements, blurriness.

We used examples like the above to train machine learning models to recognize when the camera was inside a pocket or purse (above, left), or when a finger or hand was in front of the lens (above, right). While it wasn't immediately intuitive to train models to ignore things, over time it became a crucial strategic piece in our design. By ruling out the stuff the camera wouldn't need to waste energy processing (because no one would find value in it), the overall baseline quality of captured clips rose significantly.

Composition

We needed to train models about stability, sharpness, and framing. Without careful attention, a face detection model will appreciate a face at the edge of the frame just as much as one in the center.

In an effort to train a model about subject continuity, it was important to selectively highlight examples where a subject was consistently well-framed (such as above, left).

Social norms

Familiarity is such a cornerstone of photography. You point a camera at someone and they offer implicit consent by smiling or posing. Moreover, you're the one looking through the viewfinder framing and composing the shot. With an autonomous camera, we had to be extremely clear on who is *actually* familiar to you based on social cues like the amount of time spent with them and how consistently they've been in the frame.

Editing

Diversity and redundancy is something we take for granted in the way we shoot photos; there's a little voice in the back of our head saying, "You haven't seen anything like this!" Or, "You've got enough shots of your kid for now, relax." But our models needed a lot of help.

We approached diversity along three different vectors:

- **Time**: The simple value of time passing is an important signal to appreciate. Don't go too long without capturing something.
- **Visual**: Subtle or dramatic changes in color can tell a lot about changes in environment and activity. Try to capture moments that have distinct aesthetic qualities.
- **People**: Are you in a big group or a small group or alone? Understanding how many different familiar faces you're encountering is a crucial part of feeling like you haven't missed important moments.

Trust and self-efficacy

One of the reasons we invested in Clips was because of how deeply important it was to demonstrate the importance of on-device and privacy-preserving machine learning to the world—not to mention its remarkable capabilities (e.g. it uses less power, which means devices don't get as hot, and the processing can happen quickly and reliably without needing an internet connection). A camera is a very personal object, and we've worked hard to ensure it—the hardware, the intelligence, and the content—ultimately belongs to you and you alone. Which is why everything—and I mean everything—stays on the camera until the user says otherwise.

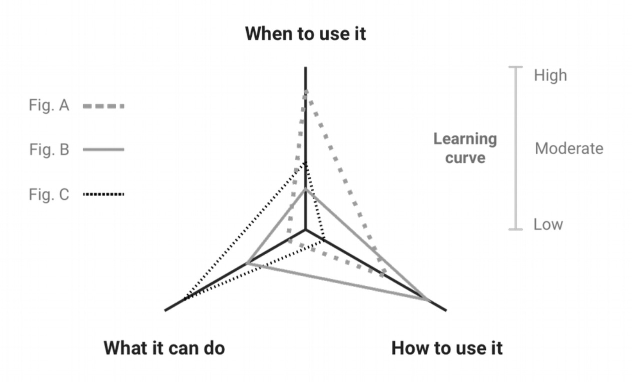
Concept budgeting

With an eye on trust and self-efficacy, we were also very intentional in the way we approached UI design. At the start of the project, that meant working through a few of our own funny assumptions about how "out-there" an AI-powered product needed to be.

When we reach into our brains for future-tech reference points, many designers will jump to the types of immersive experiences seen in movies like *Minority Report* and *Blade Runner*. But just imagine of how crazy it'd be to actually explain something like the UI in Minority Report to users: *Here, just extend your arm out, wait two seconds, grasp at thin air, then fling wildly to the right while rotating your hand counter-clockwise. It's easy!* Almost every sci-fi faux UI is guilty of something similar; as if the complexity of an interaction model needs to keep pace with the complexity of the system it's driving. But that's sort of where we were for awhile during our early design phase, and we got away with it in large part for three reasons:

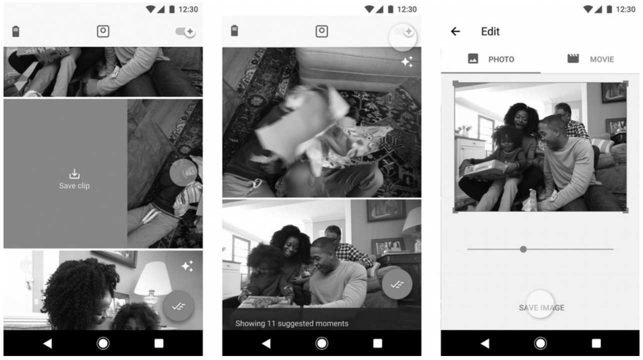
- We were showing people fake content in an obviously simulated environment, where they had no real connection to the imagery. Note that this issue isn't unique to AI; it's often one of the confounding factors when you bring people into the usability lab.
- We were surrounded by people every day who were all speaking the same language; thinking deep thoughts about AI-enabled futures. We were making the mistake of losing touch with the reference points that everyone else would bring to the table.
- We thought our new designs were super cool, so we gave ourselves a healthy amount of forgiveness when people didn't immediately get it.

Over time, we snapped out of it. We began fiercely reducing complexity in the UI, and made *control* and *familiarity* cornerstones of our experiential framework. We added a software viewfinder and a hardware capture button to the camera. We made sure that the user had the final say in curation; from the best still frame within a clip to its ideal duration. And we showed users more moments than what we necessarily thought was *just right*, because by allowing them to look a bit below the 'water line' and delete stuff they didn't want, they actually developed a better understanding of what the camera was looking for, as well as what they could confidently expect it to capture in the future.



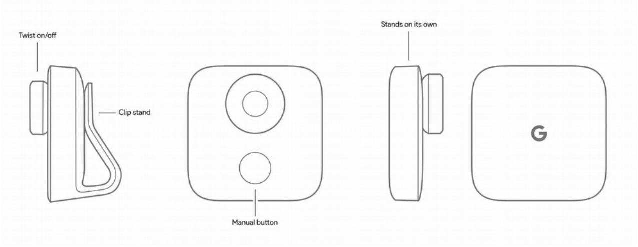
Most products have at least some learning curve, but with the added overhead of AI hype, it's especially important to 'spend' wisely on your user's cognitive load. When the context of use is novel to the user [figure A], bias for dependability. When there are a lot of new UI tricks to learn [figure B], make sure the primary use cases are super relatable. And when the functionality of the product is especially dynamic [figure C] , your UI should be flush with familiar patterns.

Through this process we discovered another critically important finding for testing an AI-powered product: fake it till you make it. If forced to choose, it's leaps-and-bounds more useful to prototype your UX with a user's real content using a Wizard of Oz approach than it is to test with real ML models. The latter takes an incredibly long time to build and instrument (and is far less agile or adaptive than traditional software development, so it's more costly to swing and miss), while the former affords you genuine insights into the way people will derive value and utility from your (theoretical) product.



Users preview their clips by streaming them from the camera. On the far left, users choose which clips they want saved to their phone. In the middle, users can toggle on a "suggested" view. On the right, users can pinpoint the exact frame they want to save as a still photo.

In the context of subjectivity and personalization, perfection simply isn't possible, and it really shouldn't even be a goal. Unlike traditional software development, ML systems will never be "bug-free"—insofar as a bug is defined as something that prevents the user from arriving at a specific linear outcome—because prediction is an innately fuzzy science. But it's precisely this fuzziness that makes ML so useful! It's what helps us craft dramatically more robust and dynamic 'if' statements, where we can design something to the effect of "when something looks sort of like x, do y." And in that departure from rigid logic rules, we also needed to depart from traditional forms of measuring engagement. Success with Clips isn't just about keeps, deletes, clicks, and edits (though those are important), it's about authorship, co-learning, and adaptation over time. We really hope users go out and play with it.



The camera turns on and off with simple twist of the lens and has a shutter button on the front for manual capture.

Designing with purpose

By re-orienting the conventional AI paradigm from finding ways to make the machine smarter, to exploring ways to augment human capability, we can unlock far greater potential in machine learning. It can become a tool for unprecedented exploration and innovation; a tool to help us seek out patterns in ourselves and the world around us. As human-centered practitioners, we have a tremendous opportunity to shape a more humanist and inclusive world in concert with AI, and it starts by remembering our roots: finding and addressing human needs through observation and experimentation, upholding humane values, and designing for augmentation[2], not automation.

The role of AI shouldn't be to find the needle in the haystack for us, but to show us how much hay it can clear so we can better see the needle ourselves.

[2] Augmenting Human Intellect: A Conceptual Framework, Engelbart 1962