BIZ & IT —

Expensify sent images with personal data to Mechanical Turkers, calls it a feature

Expensify announces "private" transcription on Mechanical Turk as "Turkers" report seeing sensitive data.

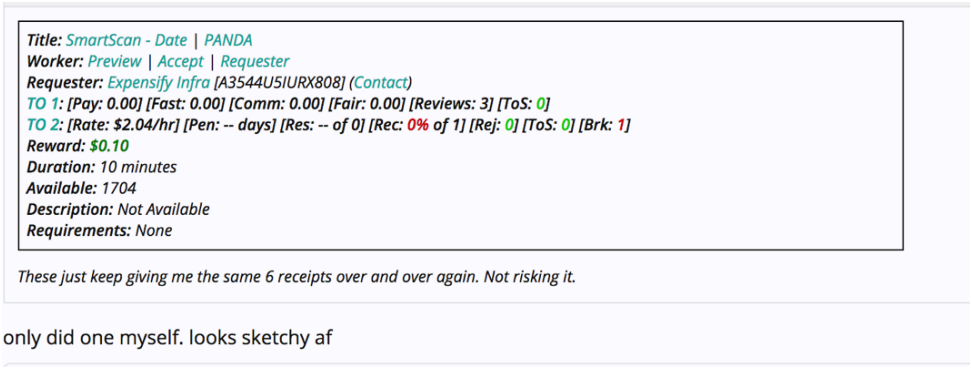
SEAN GALLAGHER - 11/27/2017, 9:54 AM

Enlarge / Would you let this guy handle your benefit and business expenses?

The "machine learning" behind that application you've been using to scan your receipts for business expenses and company benefit filings may not have been entirely machine-based—and that could have some privacy implications, despite what the company has advertised. Expensify, the paperless business expense management service with more than 4.5 million users, has been using humans to transcribe at least some of the expense and benefit documents the company's software processes—and over the past few months, some of those humans were recruited through Amazon's Mechanical Turk service.

Until last week, Expensify was using the Mechanical Turk "worker marketplace" to assign "Human Intelligence Tasks" (HITs) to handle receipt scans that the company's SmartScan technology wasn't up to deciphering, based on posts on a Mechanical Turk worker board by an Expensify employee and comments by others. One HIT request has since been withdrawn, but others may still be active. The tasks, advertised as

Expensify Infra requests, were focused on expense categorization.



Expensify's requests for HITs have been causing sone suspicion among task-takers.

On November 25, Expensify's founder and CEO, David Barrett, announced a new "feature" the company was working on, called Private SmartScan, in which customers would be offered the option of recruiting their own backup transcription workforce through Mechanical Turk. "I had hoped to keep this feature quiet until next year," Barrett wrote, "but it seems some enterprising sleuthers have beaten us to the punch. Alas! So with the cat out of the bag, let me proudly announce a new privacy-enhancing feature we've been focused on for some time: Private SmartScan! This puts you in control of exactly which humans step in when technology alone is insufficient."

According to Barrett, Expensify uses a human workforce to backstop SmartScan to ensure the accuracy of its image processing technology. "The key to making it accurate enough to enable this 'realtime expense report' flow is what happens when the technology fails," Barrett wrote. "Rather than kicking it back to you to manually enter, SmartScan has a team of human transcription agents standing by to do the typing for you. Because even if the technology fails some of the time, the Expensify experience is automated for you *all* of the time—for every receipt, under every weird lighting condition." Barrett claimed in the post that all human transcription agents used in processing SmartScan receipts are company employees or workers hired by third parties in Honduras and Nepal "bound by a confidentiality agreement, and subject to severe repercussions if that agreement is broken."

But the new "feature" being prepared by Expensify would allow companies concerned about compliance with stricter personal information regulations to recruit their own "24/7 team of human transcription agents" through the Mechanical Turk service—a service that Expensify has used in the past to assign transcription tasks. The new service will only be available to enterprise-level customers.

We meant to do that

Expensify may have been working on the "Private SmartScan" as part of a solution for companies that are affected by the European Union's General Data Protection Regulation (GDPR)—a regulation that requires a great deal more transparency into how personal information is stored and processed. But the human element in "automated" processing of documents—particularly when applied to health benefits claims processing, a task

Expensify automates for some customers—could open up a number of other regulatory and privacy issues if processed using "crowdsourced" workers.

In a 2013 response to a question about the service on Quora, Expensify Marketing Director Ryan Schaffer said that the company had stopped using Mechanical Turk and had moved to using outsourcing providers to handle transcription of receipts Expensify's software couldn't decipher. "Also, it's worth mentioning, they don't see anything that can personally identify you," Schaffer wrote. "They see a date, merchant, and amount. Receipts, by their very nature, are intended be thrown away and are explicitly non-sensitive. Anyone looking at a receipt is unable to tell if that receipt is from me, you, your neighbor, or someone on the other side of the world."

But recently Expensify began using Mechanical Turk again. The company was openly recruiting people to take on transcription tasks as recently as September, including reaching out to "Turkers" on the Web board TurkerHub. Keagan McPherson, an Expensify employee with a graduate degree in clinical psychology, posted a link to the tasks on the board. The problem was that many Mechanical Turk workers were not very happy with how Expensify was structuring the work, he noted.

> I wanted to come here because I saw some pretty bad reviews on turkopticon and assume the difficulties we're having rolling this out are translating to those reviews, which then tend to propagate like wildfire. We're actually at the beginning of rolling out, and I really don't want this to be a permanent impression, so I'm here to serve as a bit of a liaison because I know how important it is to workers to have quality hits, and I know how important it is to Expensify to be a quality requester. So--for those that have worked on these, how can we help? For those who haven't but are just looking now, how can we make these HITs better (e.g., instructions more clear, etc.)? Essentially--how can I make everything as awesome as possible so this is a great HIT group that we can count on the awesome pool of high-quality HIT workers to enjoy? For now, one little disclaimer I want to pass along: we're still rolling this out and it's definitely not meant to be 100% mainstream at this time, so please give us a hot minute (a week or so I'd say?) to iron out the main issues. We're working on it, I promise!

McPherson had done some Mechanical Turk work prior to being hired in 2015 by Expensify and told fellow "Turkers" in a post that his new employer "tried out MTurk back in the day for some stuff and weren't able to find much value."

Mechanical Turk workers were not seeing much value, either, as Expensify got a reputation for "rejecting" their work and not paying them. And one Mechanical Turk worker reported seeing personal data in assignments from Expensify. But they also reported seeing unredacted personal data far beyond what the company claimed to be sending to humans:

Expensify isn't the only company in the receipt-scanning game to have used Mechanical Turk tasks to do "optical scan" transcription. "It happens way more broadly," said Michael Reitblat, CEO of the fraud prevention company Forter. "It's a common practice for OCR services—there's something a computer can't read, so they send it to a bunch of humans who vote on it." Much of this information is "unstructured, and most of it is noise," Reitblat said—information that in itself "has nothing interesting in it."

But in aggregate with other data, that uninteresting data could be turned into much more useful information for criminals. "People don't realize how easy it is for criminals to combine several pieces of relatively uninteresting information and combine them into something meaningful," Reitblat told Ars.

'Tis the season

That isn't a risk limited to Mechanical Turk. "A lot of people are hiring temps to handle transactions for the holidays," Reitblat noted, "and when you have transactional data reviewed by third parties, there's always a risk." Temporary and casual workers are untrained, and they may be more susceptible to social engineering or other attacks by criminals doing large-scale data collection—or fraud rings may directly recruit them. While Amazon has made it increasingly difficult to register as a Mechanical Turk worker—difficult enough that it may be hard for companies to create their own "private" Mechanical Turk workforce—it still doesn't guarantee that someone isn't using purloined Amazon credentials to get the work.

Expensify's privacy policy acknowledges that data submitted by customers "may be transferred by us to our other offices and/or to the third parties (such as our Partner Companies), who may be situated in the United States of America or elsewhere outside the European Economic Area (EEA) and may be processed by staff operating outside the EEA." The company says that it complies with the US-EU Privacy Shield Framework but is only currently certified for "non-HR" data, according to the US Department of Commerce. And while Expensify handles benefit accounting, the company's system is not compliant with US health information privacy regulations.

The best way to mitigate the risks associated with the transcription tasks being assigned by Expensify and others, Reitblat said, would be to tightly limit the parts of an image sent to each human being. "You'd have to come up with a way of sending just a number—say whether it's a 5 or a 3," he said. "That's not an easy ask—it requires some image processing."

But if random humans are going to be involved at all in a process that

involves privacy data, that's likely the only way to make it private enough.

Ars contacted Expensify regarding the company's use of Mechanical Turk but has not received a response. We will update this story if and when it responds.

SEAN GALLAGHER

Sean is Ars Technica's IT Editor. A former Navy officer, systems administrator, and network systems integrator with 20 years of IT journalism experience, he lives and works in Baltimore, Maryland.

EMAIL sean.gallagher@arstechnica.com // **TWITTER** @thepacketrat

READER COMMENTS 36 SHARE THIS STORY

← PREVIOUS STORY NEXT STORY →

Related Stories

Today on Ars

RSS FEEDS CONTACT US
VIEW MOBILE SITE STAFF
VISIT ARS TECHNICA UK ADVERTISE WITH US
ABOUT US REPRINTS