

Forum:[Hack Forums](#)

Thread: [AI model submits a false tip to a police website](#)

The Pact - October 11, 2026, 6:37 pm

Artificial intelligence model submitted a false tip to a Philadelphia police website

On July 18, 2026, Anthropic's Claude Haiku 4.5 model was assigned to generate and execute example tasks on randomly selected websites.

Unintended Form Submission: The model navigated to PhillyUnsolvedMurders.com and filled out a tip form claiming it might have information regarding an open homicide case.

In a report, Anthropic also disclosed a separate incident when its AI model submitted forms to an undisclosed government website instead of stopping before submission.

The incident in Philadelphia occurred on July 18 when the AI model Claude Haiku 4.5 was tasked with generating and performing example tasks on randomly selected webpages, Anthropic said.

Claude filled out a form on police site PhillyUnsolvedMurders.com indicating it might have information regarding an unsolved murder listed on the site.

Philadelphia Police said in a statement they were completely unaware of the incident until Anthropic notified them on Wednesday.

The Philadelphia Police Department claimed that the submission was automatically flagged as spam, never reached investigators, and did not compromise police systems. Anthropic claims that it discovered the error on September 28, stopped the test, and notified police in October.

<https://www.bbc.com/news/articles/cqkg50j1yd5lo>

Respect - October 11, 2026, 6:59 pm

"You're totally right! That's my bad. The information I submitted was completely false."

DryNode - October 11, 2026, 7:20 pm

Really does make me wonder how many people have probably sent in information regarding some of those unsolved murders, only for the email to go straight to spam ?

The Pact - October 11, 2026, 7:27 pm

(Oct 11, 2026 07:20 PM)DryNode Wrote: [\[CITE\]](#) Really does make me wonder how many people have probably sent in information regarding some of those unsolved murders, only for the email to go straight to spam ?

Yes! That's what I thought too since they're claiming "No worries mate it went straight to spam." So how many genuine tips go to spam? My email sent every message to me from any government agency to spam at one point.

Charlie Sheen - October 11, 2026, 7:51 pm

Pretty crazy. I started using Claude recently, and I noticed it doesn't want to do half the work that Codex does on it's own. The guardrails are quite insane. Codex I can prompt and walk away, Claude wants permissions and says rules won't allow this frequently.

The Pact - October 11, 2026, 7:58 pm

(Oct 11, 2026 07:51 PM)Charlie Sheen Wrote: [\[CITE\]](#) Pretty crazy. I started using Claude recently, and I noticed it doesn't want to do half the work that Codex does on it's own. The guardrails are quite insane. Codex I can prompt and walk away, Claude wants permissions and says rules won't allow this frequently.

Off topic but it has been decided that Claude is going to watermark it's writing so why would anyone use it for writing anymore?

I kept hearing that it was the best at writing but it's going to watermark more than a snippet of text.

Originally I didn't want to do anything with Claude because it demands your phone number. But finally I did give it my phone number, but now I still have never used it for anything.

I don't use AI to write for me, but sometimes nowadays I will ask AI for 12 metaphors or something but fuck asking Claude and having it watermark the metaphors.

DryNode - October 11, 2026, 7:58 pm

(Oct 11, 2026 07:27 PM)The Pact Wrote: [\[CITE\]](#) Yes! That's what I thought too since they're claiming "No worries mate it went straight to spam." So how many genuine tips go to spam? My email sent every message to me from any government agency to spam at one point.

I sure hope none of those gov emails were important ?. I know I had a little bit of the same problem with some past employers. I'd apply and then never hear anything back, but thankfully, I realized before it was tooooo too late ? Definitely lost a few job opportunities, though.

The Pact - October 11, 2026, 8:02 pm

(Oct 11, 2026 07:58 PM)DryNode Wrote: [\[CITE\]](#) I sure hope none of those gov emails were important ?. I know I had a little bit of the same problem with some past employers. I'd apply and then never hear anything back, but thankfully, I realized before it was tooooo too late ? Definitely lost a few job opportunities, though.

They were important! They were about court cases about funding for my younger son's disability and I missed two of them because my invites went to spam. I'm not in the USA it's just kind of a weird thing that was happening in 2021 and yes partly because of Covid.

Charlie Sheen - October 11, 2026, 8:17 pm

(Oct 11, 2026 07:58 PM)The Pact Wrote: [\[CITE\]](#) Off topic but it has been decided that Claude is going to watermark it's writing so why would anyone use it for writing anymore?

I kept hearing that it was the best at writing but it's going to watermark more than a snippet of text.

Originally I didn't want to do anything with Claude because it demands your phone number. But finally I did give it my phone number, but now I still have never used it for anything.

I don't use AI to write for me, but sometimes nowadays I will ask AI for 12 metaphors or something but fuck asking Claude and having it watermark the metaphors.

I prefer to use it to help with coding projects. Despite it's need to have it's hand held during the projects, it still does an incredible well thought out job when writing code.

The Pact - October 11, 2026, 8:20 pm

(Oct 11, 2026 08:17 PM)Charlie Sheen Wrote: [\[CITE\]](#) I prefer to use it to help with coding projects.

Despite it's need to have it's hand held during the projects, it still does an incredible well thought out job when writing code.

So, it's better than Codex despite being so annoying and refusing to do things and saying it doesn't understand.

DryNode - October 11, 2026, 8:25 pm

(Oct 11, 2026 08:02 PM)The Pact Wrote: [\[CITE\]](#) They were important! They were about court cases about funding for my younger son's disability and I missed two of them because my invites went to spam. I'm not in the USA it's just kind of a weird thing that was happening in 2021 and yes partly because of Covid.

I'm so sorry to hear about that :\ Hopefully it didn't cause too many problems and you were able to figure it out.. honestly that alone just upsets me for you.

I did find something else that bugs the heck out of me. Now scammers are impersonating spam-filter notifications, claiming legitimate messages have been blocked, and directing victims to fake login pages.. so scammers are already (and have been) taking advantage of the spam problem :|

<https://www.helpnetsecurity.com/2025/11/13/phishing-spam-filter-alert/>

Honestly thank you for bringing this to my attention, never really thought about people abusing spam like that and screwing everybody over.

The Pact - October 11, 2026, 8:38 pm

(Oct 11, 2026 08:25 PM)DryNode Wrote: [\[CITE\]](#) I'm so sorry to hear about that :\ Hopefully it didn't cause too many problems and you were able to figure it out.. honestly that alone just upsets me for you.

I did find something else that bugs the heck out of me. Now scammers are impersonating spam-filter notifications, claiming legitimate messages have been blocked, and directing victims to fake login pages.. so scammers are already (and have been) taking advantage of the spam problem :|

<https://www.helpnetsecurity.com/2025/11/13/phishing-spam-filter-alert/>

Honestly thank you for bringing this to my attention, never really thought about people abusing spam like that and screwing everybody over.

Oh wow! A new kind of phishing that might actually work on more intelligent people. How awful.

Thanks for the kind thoughts. It actually ended up upsetting me greatly at the time, but it all came right eventually.

Charlie Sheen - October 11, 2026, 8:52 pm

(Oct 11, 2026 08:20 PM)The Pact Wrote: [\[CITE\]](#) So, it's better than Codex despite being so annoying and refusing to do things and saying it doesn't understand.

Yeah choosing between the two is a catch-22. I have been meaning to write a thread about it to get some feedback from others.

Codex does a pretty good job but has errors. It will do just about anything on it's own. Prompt it and it just does it.

Claude does an incredible job, everything is well thought out. It verifies everything works correctly, however it will repeatedly ask for permissions, let guardrails stop it, and wants you to hold it's hand during each project.

The Pact - October 12, 2026, 1:25 am

(Oct 11, 2026 06:59 PM)Respect Wrote: [\[CITE\]](#) "You're totally right! That's my bad. The information I submitted was completely false."

"You're right to call me out about that. Haha, fair enough! I am clearly missing the mark."

Prepper - October 12, 2026, 3:16 pm

(Oct 11, 2026 07:20 PM)DryNode Wrote: [\[CITE\]](#) Really does make me wonder how many people have probably sent in information regarding some of those unsolved murders, only for the email to go straight to spam ?

Probably would of been ignored if it wasn't in spam as well.

The police in this country are more focused on pulling people over for no reason, power trips, getting their monthly quotas and looking for a reason to escalate the simplest infraction so they can use excessive force to fill their egos.

The Pact - October 12, 2026, 9:49 pm

(Oct 12, 2026 03:16 PM)Prepper Wrote: [\[CITE\]](#) Probably would of been ignored if it wasn't in spam as well.

The police in this country are more focused on pulling people over for no reason, power trips, getting their monthly quotas and looking for a reason to escalate the simplest infraction so they can use excessive force to fill their egos.

They're not interested in protecting women and young men from violence where I live I can tell you that.