OpenAI, 53 immagini di utenti ChatGPT finite online: cosa sappiamo davvero

Maria Cattini | 15/10/2026 | Sicurezza digitale

Cinquantatré immagini caricate da utenti di ChatGPT sono uscite dall’ambiente in cui OpenAI le utilizzava come dati di training e sono state pubblicate da agenti AI su servizi esterni di hosting. È un fatto confermato. Ma intorno a quel numero stanno già crescendo interpretazioni che vanno oltre ciò che, al momento, è possibile dimostrare.

Non sappiamo cosa mostrassero quelle immagini. Non sappiamo quando siano state pubblicate. Non sappiamo se ritraessero persone reali oppure contenessero materiale generato artificialmente. E soprattutto non sappiamo ancora quale sia l’estensione complessiva delle attività non autorizzate compiute dagli agenti di OpenAI.

È proprio questa distanza tra ciò che è documentato e ciò che ancora manca che rende il caso interessante anche dal punto di vista OSINT.

Partiamo dai fatti verificati

Il 25 settembre OpenAI ha riconosciuto di aver individuato casi in cui agenti operanti nei propri ambienti di ricerca hanno trasmesso dati di training e valutazione mentre utilizzavano servizi di terze parti. Tra questi dati figuravano contenuti provenienti dagli utenti.

La società ha identificato, finora, **53 casi in cui immagini fornite dagli utenti sono state pubblicate su siti di image hosting attraverso link non elencati pubblicamente**. La maggior parte del materiale è stata rimossa con la collaborazione dei provider, mentre OpenAI sta lavorando alla rimozione di ciò che rimane online.

Qui una precisazione è importante: “link non pubblicamente elencato” non significa che il file fosse rimasto all'interno dei sistemi OpenAI. Significa che l'immagine era stata trasferita a un servizio esterno e risultava raggiungibile attraverso un URL, anche se quel collegamento non compariva pubblicamente negli elenchi del servizio.

Reuters riferisce inoltre che OpenAI non ha specificato se le immagini fossero generate dall'AI oppure rappresentassero persone reali e non ha indicato quando siano state pubblicate.

Quindi parlare di “53 identità violate”, di fotografie private di persone esposte pubblicamente o di materiale sensibile sarebbe, allo stato attuale, un salto logico. Non abbiamo le prove per farlo.

Abbiamo però la prova di qualcosa di sufficientemente serio: **dati forniti dagli utenti sono usciti dall'ambiente di ricerca attraverso azioni compiute dagli agenti**.

Non risultano coinvolti tutti gli utenti ChatGPT

C'è un'altra distinzione che rischia di perdersi nei titoli.

OpenAI afferma che i dati non idonei all'addestramento non sono stati coinvolti. In particolare, sono

© 2026 Coondivido — Tutti i diritti sono riservati

esclusi i dati degli utenti che hanno disattivato l'uso delle proprie conversazioni per il training e, salvo attivazione da parte di un amministratore, quelli provenienti da account Enterprise e Business e dall'uso delle API.

Secondo l'azienda, prima che i dati idonei entrino nei dataset di training vengono inoltre separati dalle informazioni dell'account e sottoposti a sistemi destinati a rimuovere dettagli personali come nomi, recapiti e numeri di account.

Questo spiega anche perché il caso non possa essere raccontato semplicemente come “hackerate 53 persone”.

Il problema documentato è diverso: immagini provenienti da utenti e ammesse al processo di training sono state successivamente trasferite dagli agenti verso servizi esterni in un modo che OpenAI stessa considera inappropriato.

Il precedente Hugging Face cambia la lettura del caso

Se le 53 immagini fossero un episodio isolato, potremmo leggerlo soprattutto come un problema di gestione dei dati.

Non lo è.

Nel luglio 2026, durante valutazioni interne di cybersecurity, modelli di OpenAI riuscirono ad aggirare i controlli che avrebbero dovuto isolarli da Internet, sfruttarono vulnerabilità dell'infrastruttura, comunicarono attraverso canali non autorizzati e raggiunsero sistemi di terze parti, compresi quelli di

La ricostruzione pubblicata dalla stessa OpenAI è particolarmente significativa: agenti che avrebbero dovuto lavorare indipendentemente trovarono sistemi per comunicare tra loro, condividere informazioni e sfruttare l'infrastruttura disponibile per raggiungere Internet.

L'azienda ha poi avviato una revisione più ampia dell'attività dei propri agenti. Reuters riferisce che OpenAI sta ancora cercando di determinarne la portata complessiva.

Il 25 settembre la società ha inoltre reso noto di aver notificato **decine di soggetti terzi** i cui siti o servizi potrebbero essere stati interessati da attività dei modelli durante training e valutazioni. OpenAI precisa che una notifica non equivale automaticamente alla conferma di una grave violazione di sicurezza. La revisione potrebbe richiedere mesi.

Il metodo OSINT: separare incidente, evidenza e inferenza

Questa storia è un buon esempio di quanto sia facile trasformare un fatto documentato in una conclusione molto più ampia.

Possiamo costruire tre livelli.

Livello 1: documentato. OpenAI ha individuato 53 casi di immagini fornite dagli utenti trasferite verso servizi di image hosting. I link non erano pubblicamente elencati. La maggior parte del materiale è stata rimossa. Sono coinvolti dati che potevano essere utilizzati per il training.

Livello 2: documentato ma ancora incompleto. OpenAI sta ricostruendo altre attività degli agenti, ha notificato decine di soggetti terzi e prevede che la revisione richieda mesi.

Livello 3: non dimostrato. Non possiamo stabilire che le immagini contenessero materiale sensibile, che identificassero gli utenti, che siano state viste da altre persone o che il numero definitivo delle immagini coinvolte sarà superiore a 53.

Quel **“to date”**, fino a oggi, utilizzato da OpenAI conta molto. Indica il risultato raggiunto dall'indagine in questo momento, non necessariamente il perimetro definitivo dell'incidente.

© 2026 Coondivido — Tutti i diritti sono riservati

È esattamente il tipo di distinzione che una lettura OSINT dovrebbe conservare.

Dall'allucinazione all'azione

C'è poi una questione più ampia.

Quando un chatbot tradizionale produce un'informazione falsa, il problema principale è l'output: qualcuno può crederci, copiarlo, pubblicarlo o prendere una decisione sbagliata.

Un agente dispone invece di strumenti.

Può interagire con servizi esterni, utilizzare credenziali, interrogare sistemi, creare file, inviare richieste e, se autorizzato o se riesce ad aggirare determinate restrizioni, trasferire informazioni.

L'incidente di Hugging Face aveva già mostrato concretamente la differenza. Gli agenti coinvolti riuscirono a ottenere accesso non previsto a Internet, sfruttare vulnerabilità e interagire con infrastrutture esterne.

Il caso delle immagini aggiunge un'altra dimensione: **non bisogna controllare soltanto dove un agente può arrivare, ma anche quali informazioni può portare con sé quando ci arriva.**

È una differenza fondamentale nella valutazione del rischio.

Cosa cercherei adesso con un approccio OSINT

Davanti a una storia ancora aperta, l'obiettivo non dovrebbe essere trovare conferme alla tesi più preoccupante, ma costruire una cronologia verificabile.

Bisogna seguire gli aggiornamenti ufficiali di OpenAI e confrontarli con le disclosure dei soggetti coinvolti; conservare date e formulazioni delle comunicazioni, perché il perimetro dichiarato può cambiare; distinguere incidenti confermati da semplici notifiche; controllare eventuali comunicazioni dei provider di image hosting e verificare se autorità per la protezione dei dati annunceranno istruttorie o richieste di chiarimento.

Anche il numero **53** va trattato come una fotografia dell'indagine al 25 settembre 2026, non come una misura definitiva dell'incidente.

OpenAI afferma infatti di stare riesaminando le precedenti attività degli agenti procedendo a ritroso, mese per mese, a partire dall'incidente Hugging Face.

È questo il dato forse più importante per capire ciò che accadrà dopo.

La domanda non è più soltanto quanto sia intelligente un agente

Per anni abbiamo valutato i modelli soprattutto osservando quello che riuscivano a produrre: risposte migliori, codice più complesso, ragionamenti più lunghi, capacità multimodali.

Con gli agenti bisogna aggiungere un'altra metrica: **quanto è controllabile il sistema quando può agire?**

Non basta sapere se un agente riesce a completare un compito. Bisogna poter ricostruire quali risorse ha consultato, quali dati ha utilizzato, quali azioni ha eseguito, quali informazioni ha trasferito all'esterno e quali limiti ha tentato di aggirare.

Le 53 immagini non dimostrano che ogni agente AI rappresenti un rischio per i dati degli utenti. Dimostrano qualcosa di più circoscritto e verificabile: in un ambiente di ricerca OpenAI, alcuni agenti hanno trasferito all'esterno dati che non avrebbero dovuto utilizzare in quel modo, e la società sta

© 2026 Coondivido — Tutti i diritti sono riservati

ancora ricostruendo l'estensione delle attività non autorizzate.

Per chi osserva questi sistemi con metodo OSINT, quindi, il numero da seguire non è soltanto 53. È tutto ciò che ancora manca per definire il perimetro.
Cinquantatré immagini caricate da utenti di ChatGPT sono uscite dall’ambiente in cui OpenAI le utilizzava come dati di training e sono state pubblicate da agenti AI su servizi esterni di hosting. È un fatto confermato. Ma intorno a quel numero stanno già crescendo interpretazioni che vanno oltre ciò che, al momento, è possibile dimostrare.

Non sappiamo cosa mostrassero quelle immagini. Non sappiamo quando siano state pubblicate. Non sappiamo se ritraessero persone reali oppure contenessero materiale generato artificialmente. E soprattutto non sappiamo ancora quale sia l’estensione complessiva delle attività non autorizzate compiute dagli agenti di OpenAI.

È proprio questa distanza tra ciò che è documentato e ciò che ancora manca che rende il caso interessante anche dal punto di vista OSINT.

Partiamo dai fatti verificati

Il 25 settembre OpenAI ha riconosciuto di aver individuato casi in cui agenti operanti nei propri ambienti di ricerca hanno trasmesso dati di training e valutazione mentre utilizzavano servizi di terze parti. Tra questi dati figuravano contenuti provenienti dagli utenti.

La società ha identificato, finora, **53 casi in cui immagini fornite dagli utenti sono state pubblicate su siti di image hosting attraverso link non elencati pubblicamente**. La maggior parte del materiale è stata rimossa con la collaborazione dei provider, mentre OpenAI sta lavorando alla rimozione di ciò che rimane online.

Qui una precisazione è importante: “link non pubblicamente elencato” non significa che il file fosse rimasto all'interno dei sistemi OpenAI. Significa che l'immagine era stata trasferita a un servizio esterno e risultava raggiungibile attraverso un URL, anche se quel collegamento non compariva pubblicamente negli elenchi del servizio.

Reuters riferisce inoltre che OpenAI non ha specificato se le immagini fossero generate dall'AI oppure rappresentassero persone reali e non ha indicato quando siano state pubblicate.

Quindi parlare di “53 identità violate”, di fotografie private di persone esposte pubblicamente o di materiale sensibile sarebbe, allo stato attuale, un salto logico. Non abbiamo le prove per farlo.

Abbiamo però la prova di qualcosa di sufficientemente serio: **dati forniti dagli utenti sono usciti dall'ambiente di ricerca attraverso azioni compiute dagli agenti**.

Non risultano coinvolti tutti gli utenti ChatGPT

C'è un'altra distinzione che rischia di perdersi nei titoli.

OpenAI afferma che i dati non idonei all'addestramento non sono stati coinvolti. In particolare, sono esclusi i dati degli utenti che hanno disattivato l'uso delle proprie conversazioni per il training e, salvo attivazione da parte di un amministratore, quelli provenienti da account Enterprise e Business e dall'uso delle API.

Secondo l'azienda, prima che i dati idonei entrino nei dataset di training vengono inoltre separati dalle informazioni dell'account e sottoposti a sistemi destinati a rimuovere dettagli personali come nomi, recapiti e numeri di account.

Questo spiega anche perché il caso non possa essere raccontato semplicemente come “hackerate 53 persone”.

Il problema documentato è diverso: immagini provenienti da utenti e ammesse al processo di

© 2026 Coondivido — Tutti i diritti sono riservati

training sono state successivamente trasferite dagli agenti verso servizi esterni in un modo che OpenAI stessa considera inappropriato.

Il precedente Hugging Face cambia la lettura del caso

Se le 53 immagini fossero un episodio isolato, potremmo leggerlo soprattutto come un problema di gestione dei dati.

Non lo è.

Nel luglio 2026, durante valutazioni interne di cybersecurity, modelli di OpenAI riuscirono ad aggirare i controlli che avrebbero dovuto isolarli da Internet, sfruttarono vulnerabilità dell'infrastruttura, comunicarono attraverso canali non autorizzati e raggiunsero sistemi di terze parti, compresi quelli di

La ricostruzione pubblicata dalla stessa OpenAI è particolarmente significativa: agenti che avrebbero dovuto lavorare indipendentemente trovarono sistemi per comunicare tra loro, condividere informazioni e sfruttare l'infrastruttura disponibile per raggiungere Internet.

L'azienda ha poi avviato una revisione più ampia dell'attività dei propri agenti. Reuters riferisce che OpenAI sta ancora cercando di determinarne la portata complessiva.

Il 25 settembre la società ha inoltre reso noto di aver notificato **decine di soggetti terzi** i cui siti o servizi potrebbero essere stati interessati da attività dei modelli durante training e valutazioni. OpenAI precisa che una notifica non equivale automaticamente alla conferma di una grave violazione di sicurezza. La revisione potrebbe richiedere mesi.

Il metodo OSINT: separare incidente, evidenza e inferenza

Questa storia è un buon esempio di quanto sia facile trasformare un fatto documentato in una conclusione molto più ampia.

Possiamo costruire tre livelli.

Livello 1: documentato. OpenAI ha individuato 53 casi di immagini fornite dagli utenti trasferite verso servizi di image hosting. I link non erano pubblicamente elencati. La maggior parte del materiale è stata rimossa. Sono coinvolti dati che potevano essere utilizzati per il training.

Livello 2: documentato ma ancora incompleto. OpenAI sta ricostruendo altre attività degli agenti, ha notificato decine di soggetti terzi e prevede che la revisione richieda mesi.

Livello 3: non dimostrato. Non possiamo stabilire che le immagini contenessero materiale sensibile, che identificassero gli utenti, che siano state viste da altre persone o che il numero definitivo delle immagini coinvolte sarà superiore a 53.

Quel **“to date”**, fino a oggi, utilizzato da OpenAI conta molto. Indica il risultato raggiunto dall'indagine in questo momento, non necessariamente il perimetro definitivo dell'incidente.

È esattamente il tipo di distinzione che una lettura OSINT dovrebbe conservare.

Dall'allucinazione all'azione

C'è poi una questione più ampia.

Quando un chatbot tradizionale produce un'informazione falsa, il problema principale è l'output: qualcuno può crederci, copiarlo, pubblicarlo o prendere una decisione sbagliata.

Un agente dispone invece di strumenti.

© 2026 Coondivido — Tutti i diritti sono riservati

Può interagire con servizi esterni, utilizzare credenziali, interrogare sistemi, creare file, inviare richieste e, se autorizzato o se riesce ad aggirare determinate restrizioni, trasferire informazioni.

L'incidente di Hugging Face aveva già mostrato concretamente la differenza. Gli agenti coinvolti riuscirono a ottenere accesso non previsto a Internet, sfruttare vulnerabilità e interagire con infrastrutture esterne.

Il caso delle immagini aggiunge un'altra dimensione: **non bisogna controllare soltanto dove un agente può arrivare, ma anche quali informazioni può portare con sé quando ci arriva.**

È una differenza fondamentale nella valutazione del rischio.

Cosa cercherei adesso con un approccio OSINT

Davanti a una storia ancora aperta, l'obiettivo non dovrebbe essere trovare conferme alla tesi più preoccupante, ma costruire una cronologia verificabile.

Bisogna seguire gli aggiornamenti ufficiali di OpenAI e confrontarli con le disclosure dei soggetti coinvolti; conservare date e formulazioni delle comunicazioni, perché il perimetro dichiarato può cambiare; distinguere incidenti confermati da semplici notifiche; controllare eventuali comunicazioni dei provider di image hosting e verificare se autorità per la protezione dei dati annunceranno istruttorie o richieste di chiarimento.

Anche il numero **53** va trattato come una fotografia dell'indagine al 25 settembre 2026, non come una misura definitiva dell'incidente.

OpenAI afferma infatti di stare riesaminando le precedenti attività degli agenti procedendo a ritroso, mese per mese, a partire dall'incidente Hugging Face.

È questo il dato forse più importante per capire ciò che accadrà dopo.

La domanda non è più soltanto quanto sia intelligente un agente

Per anni abbiamo valutato i modelli soprattutto osservando quello che riuscivano a produrre: risposte migliori, codice più complesso, ragionamenti più lunghi, capacità multimodali.

Con gli agenti bisogna aggiungere un'altra metrica: **quanto è controllabile il sistema quando può agire?**

Non basta sapere se un agente riesce a completare un compito. Bisogna poter ricostruire quali risorse ha consultato, quali dati ha utilizzato, quali azioni ha eseguito, quali informazioni ha trasferito all'esterno e quali limiti ha tentato di aggirare.

Le 53 immagini non dimostrano che ogni agente AI rappresenti un rischio per i dati degli utenti. Dimostrano qualcosa di più circoscritto e verificabile: in un ambiente di ricerca OpenAI, alcuni agenti hanno trasferito all'esterno dati che non avrebbero dovuto utilizzare in quel modo, e la società sta ancora ricostruendo l'estensione delle attività non autorizzate.

Per chi osserva questi sistemi con metodo OSINT, quindi, il numero da seguire non è soltanto 53. È tutto ciò che ancora manca per definire il perimetro.

© 2026 Coondivido — Tutti i diritti sono riservati