Silicon Valley, l'algoritmo anti-Isis

Facebook, Twitter, YouTube, Microsoft creano un database comune per prendere le «impronte dell'odio». Su pressione di Washington e Bruxelles

In gergo lo chiamano il super algoritmo anti Isis. In pratica si tratta di un database comune che mette in condivisione le impronte dei contenuti web riconducibili a gruppi terroristici.

L'annuncio è arrivato dritto dalla Silicon Valley. «Abbiamo stretto un accordo per combattere l'estremismo in rete», hanno spiegato Facebook, Twitter, Microsoft e YouTube in un comunicato congiunto, ultimo passo di un processo tortuoso e impervio per fermare la radicalizzazione di migliaia di giovani, spinti nelle maglie di Isis da video, immagini e proclami. Sembrano passati secoli da quando Obama, dopo la morte di Jim Foley e la diffusione del video della sua decapitazione, nell'agosto 2014 sollecitò un impegno dei big del tech nella lotta alla radicalizzazione.

Da allora tanta melma è passata sotto i ponti dei social. Bufale, insulti, contenuti inappropriati, minacce e sgozzamenti. Il problema si è esteso ben oltre il Califfato. Intanto, il tira e molla per la guerra alla jihad 3.0 tra Washington e Ue da una parte e la Silicon Valley dall'altra andava avanti.

La ragion di Stato contro la privacy. La sicurezza o la libertà di parola. Per Zuckerberg e colleghi censurare ha sempre voluto dire prestarsi a critiche che, puntuali, sono arrivate anche ieri, con le organizzazioni per la difesa della libertà del web sul piede di guerra. «Nessun contenuto viene rimosso in modo automatico», si sono affrettati a specificare i portavoce delle aziende. Non a caso per convincere i privati ad azioni concrete c'è voluta la commissaria europea Vera Jourova che domenica, in un'intervista al *Financial Times*, ha bacchettato i colossi accusandoli di non rispettare l'impegno preso con l'Ue che prevede una risposta alle segnalazioni dei contenuti di *hate speech* entro le 24 ore.

Al di là delle parole di circostanza, resta fermo il problema economico. Individuare il cyber-odio solo grazie all'occhio umano è impensabile. Servono soldi. Per qualche tempo nella Silicon ognuno ha fatto da sé. Ma più o meno tutti hanno sempre usato un sistema chiamato PhotoDna, utilizzato nella lotta alla pedopornografia, prodotto da Microsoft. In pratica si tratta di riconoscere degli schemi ricorrenti. Al posto delle foto dei bambini, le tute arancioni dei prigionieri di Isis, tolte dalla circolazione, mentre gli strateghi del Califfato replicava cambiando il colore delle divise agli ostaggi. Poi, a poco tempo di distanza dalla bufera Apple-Fbi per l'attacco di San Bernardino, in pieno dibattito sulla crittografia, Obama è tornato alla carica, richiedendo un super algoritmo per mettere ko l'Isis. Le risposte ancora una volta sono state tiepide. Ma non sono mancate. Oltre alla rimozione dei contenuti, a Mountain View, ad esempio, hanno deciso anche di fare qualcosa sul fronte della narrativa alternativa con Redirect Method. Alla divisione Jigsaw di Alphabet hanno iniziato a proporre link anti Isis a chi va a caccia di odio in rete. Anche dal mondo del no profit non sono stati a guardare. Hany Farid, uno dei padri del PhotoDna, ha adattato il sistema ai contenuti violenti e lo ha messo a disposizione dell'organizzazione Counter Extremism Project. Tutte belle idee. Ma sul fronte della rimozione per tanto tempo ciascuno ha fatto da sé. Fino a ieri, con l'annuncio del database degli *hashes*, le impronte digitali dell'odio, che — promettono i colossi— sarà attivo a partire dal 2017, e verrà presto presentato.

Difficile però cantare vittoria. Nonostante le difficoltà sul campo militare e l'uccisione di figure chiave come il portavoce Al Adnani, la propaganda dell'Isis non si ferma. Semplicemente si sposta in luoghi meno frequentati ma non per questo meno accessibili. Il tutto mentre restano aperte due domande: chi decide che cos'è «terrorismo»? E soprattutto chi stabilisce quando un contenuto va rimosso? Ora la palla è nel campo della Silicon Valley. Da vedere quanto ci resterà.

Marta Serafini
@martaserafini

La parola

HATE SPEECH

L'espressione, tradotta con «incitamento all'odio», identifica una categoria della giurisprudenza
Usa utilizzata per parole e discorsi che esprimono intolleranza per una persona o un gruppo, e che possono provocare reazioni violente