臉書演算法是如何界定仇恨言論？

研究發展部資深研究員何國華　2017 年 6 月

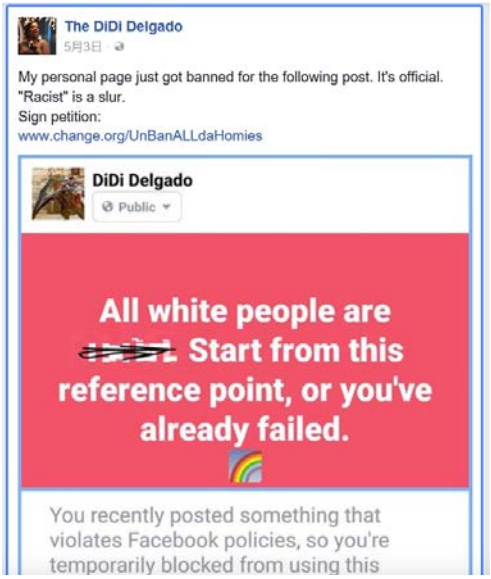
臉書成立 13 年，每月活躍用戶數剛突破 20 億人的同時，也面臨到不少的頭痛問題，尤其是全球使用者和各國政府，對於臉書假新聞和仇恨言論的批評聲浪越來越高漲，臉書爲此採取了各項監控和防堵假新聞和仇恨言論的措施。

ProPublica 資深調查記者 Julia Angwin 和 Hannes Grassegger 於 6 月 28 日刊出最新調查報導。兩人針對臉書利用演算法「秘密審查機制」(Secret Censorship Rules)，如何避免仇恨言論所做的深度採訪報導結論，直指「臉書的努力還是不夠」。

臉書演算法的設計初衷是要公平保障所有的族群和性別，但是根據臉書外流的內部文件顯示，臉書試圖利用演算法區分合法政治表達和仇恨言論兩者的差異性似乎仍力有未逮，ProPublica 報導以「臉書只保護得了白人免受傷害，卻保護不了黑人兒童」標題，嘲諷式評價臉書演算法根本是爲德不卒。

根據臉書所發布的社群準則，其中針對仇恨言論進行規範，提出「禁止基於實質上或認知上的種族、國籍、宗教、性別或性別認同、性傾向、身心障礙或疾病，而攻擊他人的內容。但是，我們允許帶有明顯幽默或諷刺的意圖，儘管這些意圖有可能被視爲潛在威脅或攻擊。這包括很多人認爲是低俗趣味的內容（如笑話、滑稽喜劇及流行歌詞等等）。」

當 6 月初倫敦發生恐攻事件時，美國國會議員 Clay Higgins 隨即在臉書上 Po 文（下圖左），公開呼籲「殺光這些極端的穆斯林份子」，臉書並未處理此篇 Po 文，現在進入 Higgins 的臉書仍可看到這篇文章。但是臉書對待波士頓詩人 Didi Delgado 在 5 月的 Po 文（下圖右），則是採取不同的做法，因爲 Delgado 指所有的白人都是種族主義者，她的 Po 文立刻遭到移除，同時帳號也被停權七天，Delgado 爲此表達抗議。



這究竟是怎麼一回事？臉書的刪文標準究竟是什麼？

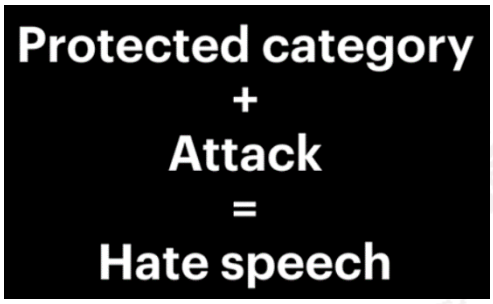
大家可能會認爲 Higgins 的文章就是仇恨言論呀，而且也不符合臉書針對仇恨言論所做的規範和定義，但是審視最後的結果，爲什麼會有這樣的差別待遇? 亦即 Higgins 的文章保留了下來，Delgado 文章遭到移除呢？真的是看不懂臉書演算法的邏輯？

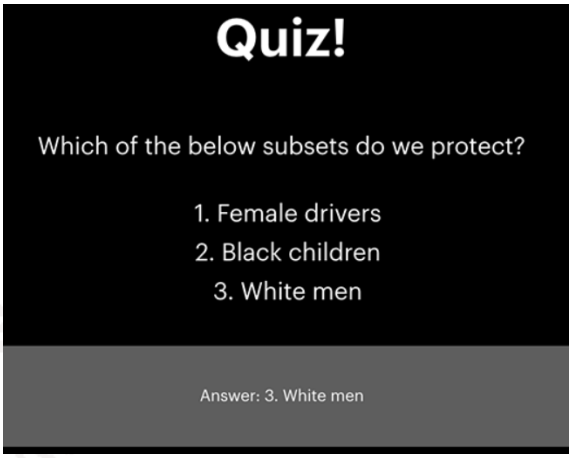
根據臉書外流文件指出，兩者差別在於：Higgins 文章是針對「極端」的特定穆斯林，Delgado 文章則是針對「所有」的白人，所以一個是針對部分人，一個是針對所有人，最後的結果就會大不相同。

報導中也披露臉書演算法標註三組人當中，究竟那一組最應受到仇恨言論的保護？三個選項包括：「女性駕駛」、「黑人兒童」、「白人男性」。

正確的答案竟然是「白人男性」。

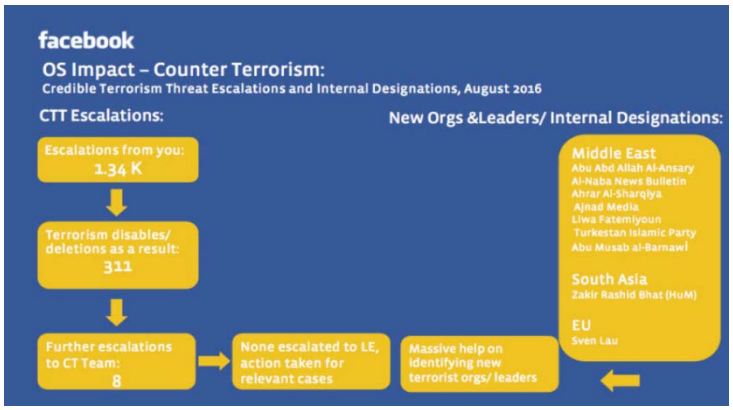
爲什麼是「白人男性」？而不是「女性駕駛」、「黑人兒童」？





原因是臉書演算法只針對種族、性議題、性取向、性別認同、信仰、血統、種族、重殘等受保護類別做爲演算基礎，觸及這些標準的 Po 文才會將涉及詛咒、誹謗、呼籲暴力等攻擊言論刪除。

因爲臉書演算法給予受保護類別的「子集」（subset）不同的定義範圍，所以「白人男性」會被視爲一個整體，必須要被保護，「女性駕駛」、「黑人兒童」，或是「極端穆斯林份子」，都是屬於臉書演算法當中整體的一個「子集」，所以才會產生令人無法理解的差異結果。



以臉書反恐 Po 文爲例，根據臉書「反恐小組」（counter-terrorism team，上圖）

的處理 Po 文流程，「反恐小組」曾經有過在一個月內標註出超過 1,300 則恐怖威脅的 Po 文，但是當主事者深陷在臉書海量的資訊當中，要如何處理這些 Po 文，根本就是一項不可能的任務，因爲 1,300 則是已經察覺有問題的 Po 文，必須處理移除，但是沒有看見的 Po 文有多少呢？可能都會是埋藏在其中的未爆彈。

尤其當不少 Po 文都採取新聞發布形式呈現時，標題是採取中性文字陳述，對此，演算法基本上是難以察覺和移除的，例如一件流血衝突的爭議性事件 Po 上臉書，採取事實新聞呈現時，最終結果可能就是 Po 文繼續留下來。

所以善用臉書進行假新聞和仇恨言論操作的人，可能都已發展出一套方法可以欺瞞過臉書演算法，這樣的結果可能會讓人很傻眼。這也是何以各國政治人物會群起痛批社群媒體可恥，死要賺取高額利潤，卻無法保障使用者不受威脅的基本要求，因此各國已研商祭出法條，大幅提高裁罰金額。

目前全球網路使用者約有 34 億人，總用戶數排在臉書之後的網路平台，包括 YouTube（15 億）、WeChat（8.89 億）、Instagram（7 億）、Twitter（3.28 億）與 Snapchat（2.55 億）。另外，臉書旗下的兩大平台 WhatsApp 與 Messenger，也都各自擁有 12 億用戶。當臉書旗下用戶數愈來愈多，平台上話題影響的人數愈來愈多，臉書對世界的影響力也愈來愈大。像是假新聞猖獗間接導致川普選上總統、暴力事件透過臉書 Live 直播等事件，都是臉書必須持續面對和處理的重要議題。

參考資料

成立 13 年，Facebook 用戶數正式突破 20 億！
https://www.bnext.com.tw/article/45104/facebook-maus-surpasses-2-billion
Facebook’ s Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children
https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms
Facebook struggles with 'mission impossible' to stop online extremism
https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossi

ble-to-stop-online-extremism