社群媒體 你相信嗎？ 網路訊息 誰說了算？

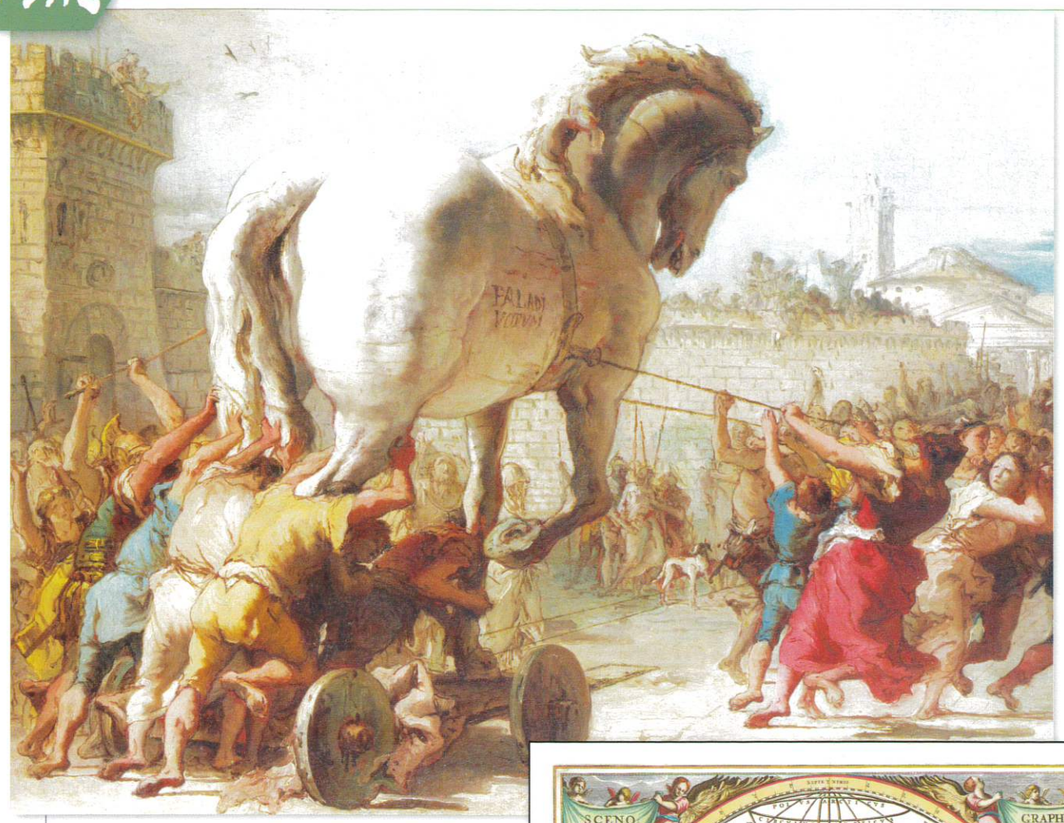
◆ 社團法人台灣 E 化資安分析管理協會、東海大學資工系 — 賴俊鳴

隨著社群媒體盛行以及步調愈來愈快的資訊速食文化，使得民眾愈來愈難接觸到專業且客觀的資訊，「假新聞」一詞漸成為互相攻訐的工具。

訊息戰並非社群媒體獨有

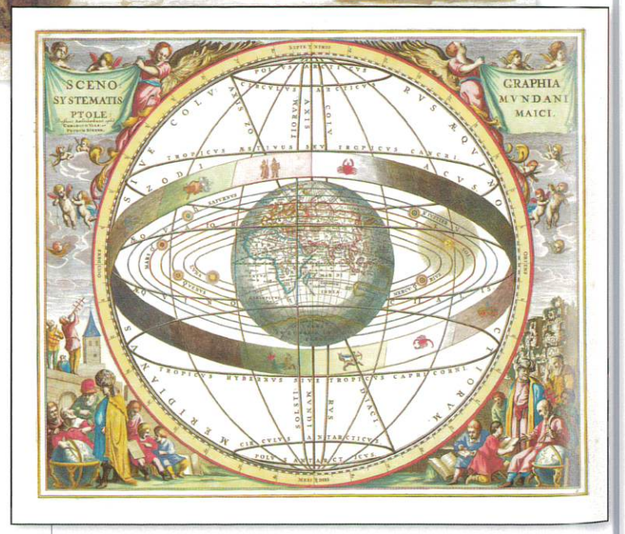
從歷史的角度來看，散播不實謠言以達到娛樂，甚至政治、經濟的目的案例比比皆是。最著名的有出自荷馬史詩與希臘神話的「特洛伊之戰」，希臘聯軍利用巨大木馬躲藏伏兵，並派奸細至特洛伊城內散布希臘聯軍已撤退，且巨大木馬為獻給神的戰利品等錯誤資訊，最終讓久攻不下的特洛伊城陷落。另一方面，訊息真偽也可能隨著時間而改變，例如：「地球繞著太陽轉」於中世紀被認為是偽科學，因該時代人們普遍相信地球乃是宇宙中心。試想，**若連最嚴謹的同儕審查之學術論文也無法保證其真確性，更遑論缺乏監管機制且幾乎人人可發文的社群媒體？**

自從美國前總統川普多次稱呼對他負面報導以及評論的資訊為「假新聞」，加上各大網路巨擘因資料隱私問題紛紛被傳喚至美國國會聽證會後，社群媒體安全與隱私議題迅速席捲全球。



歷史上著名的「特洛伊之戰」即是希臘聯軍利用巨大木馬躲藏伏兵，故意傳播錯誤訊息，成功欺瞞敵軍，最終順利奪下特洛伊城。（Photo Credit: Created by Giovanni Domenico Tiepolo, circa 1760）

臉書（Facebook）於 2018 年開始終止其他應用程式透過 API 取得帳號階層的資訊，包括討論、留言、按讚，以作為對於「劍橋分析」事件[1]的防火牆，**因為「劍橋分析」事件已證明廣告足以影響選舉結果**。[2] 推特（Twitter）更成立了一整個部門，透過人工智慧以及資料探勘技術，定期公布其檢測之可疑帳號，包括中國大陸、俄羅斯、烏干達以及委內瑞拉等國家的可疑帳號。[3]



「天動說」是一種天文學學說，認為地球是宇宙中心，其他日月星辰環繞著地球運行，初期頗為人民普遍接受；文藝復興時代後，隨著科學技術的進步，以太陽為宇宙中心的「地動說」證據逐漸出現，偽科學因此被證實取代。（Photo Credit: Created by Andreas Cellarius, 1660）

[1] Cambridge Analytica 證明透過單一平臺的帳號歷史資料之搜集、探勘與分析，足以使政治廣告公司推薦用戶相關資訊，進而影響選舉。

[2] 劍橋分析是一家英國的數據公司，創立於 2013 年，曾在 Facebook 上推出一款免費心理測驗 App，被發現在未經用戶許可的情況下，盜用 Facebook 5 千萬用戶個資，同時，也被質疑是 2016 年美國總統大選川普團隊用來左右選舉的幕後黑手。https://www.bnext.com.tw/article/55756/cambridge-analytica-election-taiwan-facebook。

[3] https://blog.twitter.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed.

「劍橋分析」是一家數據公司，曾受聘於川普競選團隊，當時在臉書推出一款心理測驗 App，取得部分用戶資訊，後被發現未經許可盜用臉書 5 千萬用戶個資，涉嫌用來操縱美國總統選情。（Photo Credit: Book Catalog, https://flic.kr/p/Hoh2iY; Gage Skidmore, https://flic.kr/p/MQVjMb）

臉書前資安長 Alex Stamos 教授將資訊頻道依據溝通方式分成數類，[4] 如圖 1 所示。倒三角形最下方為一對一的溝通模式，例如臉書即時訊息（Facebook Messenger）、Line 帳戶對帳戶等皆為此類，愈往上觸及受眾愈多，即擴大效應愈明顯，另一方面，愈往下對隱私的考量愈重。從最下方一對一的溝通模式往上依序為群組訊息（例如 Line 群組）、私有個人帳號、邀請制的粉絲專頁、公開個人帳號、公開粉絲專頁，接著透過推薦引擎將所有資訊與演算法結合，推送最容易吸引使用者花費更多時間的內容。

有創意的攻擊者帳號

相比於傳統媒體，社群媒體通常不具編輯審查機制，好處為資訊的流通更即時且開放，缺點則為資訊真偽難以驗證。

為吸引相對多的使用者透過按讚、留言、分享方式與文章互動，一個最通用且常見的特徵為發布之初，**攻擊者先利用程式同步創造一群一群的帳號**，[5] **等到收到指示要炒熱某文章後，再透過「養好的」帳號群（Clusters）向社群推薦系統發起攻擊**，結合所謂「心理學認知攻擊手法」，

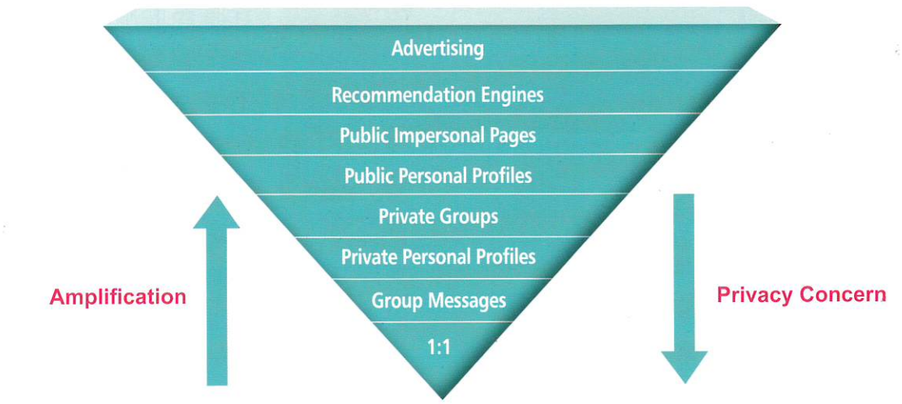


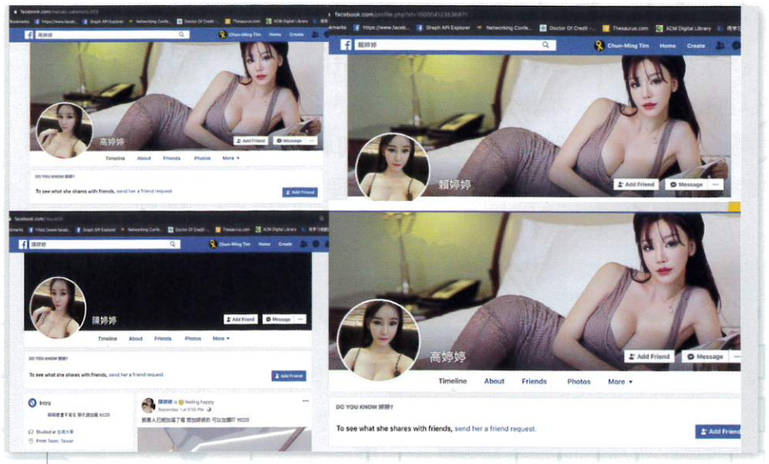
圖 1　資訊頻道分層圖

[4] 2019 年於全球頂尖資安會議 USENIX 大會分享其於臉書以及史丹佛大學長期研究社群媒體安全議題。

[5] 攻擊者可透過逆向工程方式得知如何使（欲散布）訊息被社群平臺之推薦引擎所擴展。

例如從眾效應（Bandwagon Effect）、[6] 沉默螺旋（Spiral of Silence），[7] 使某文章收到比預期強大的效果，進而影響使用者現實社會中的行為。

根據筆者研究，這些殭屍帳號取名相當有創意，例如有古詩詞家的帳號群（李清照、李白、杜甫等）、古代皇帝的帳號群（李世民、朱元璋、朱標等），還有類似 AI 影像處理變換的美女身分，**創立不同帳號，彼此為朋友，經常針對某議題「共同出擊」**，因此，偵測「同步非真實行為」（Coordinated Inauthentic Behavior, CIB）議題目前為社群媒體安全之重點發展領域。



殭屍帳號通常採用某種固定模式設定身分，圖為 4 個不同帳號，但影像資料皆相同，彼此為朋友，且經常針對某一議題「共同出擊」。（圖片來源：作者提供）

然而刪除帳號此一舉動，對於投資人衡量社群公司有著負面影響，例如 2022 年 Q1 因為臉書月活躍用戶（Monthly Active Users, MAU）數量首度下滑，造成一個禮拜股價大跌 39%，市值蒸發約 7.4 兆臺幣；[8] 因此各大社群公司對於處理此類可疑帳號仍然保持曖昧不明的態度。

驗證訊息真偽方式

目前學術界以及社群媒體公司對於社群平臺上的資安問題大略分成四個方向來偵測與緩解，如圖 2 所示。最上層為根據文字語義與內容，通常透過大量人力來逐一檢驗訊息的真偽，以類似學術論文之同儕審核機制，根據每一則回報訊息發表核實報告。國內常見的組織有「台灣事實查核中心」、「Cofacts 真的假的」以及「MyGoPen（麥擱騙）」等；[9] 缺點為人工審核耗時費工，且資金來源若為商業以及政府計劃，易招惹不中立之批評。最底層為帳號階層的防禦，透過 Captcha 以及帳號活動關聯等技術，綁定自然人與帳號之間關係，抑制殭屍帳號的創造與維持；

6 指人們受到多數人的思想或行動影響，而跟從大眾想法或行為的現象。

7 當人們發現自己意見與主流意見不同時，因為害怕被孤立或迫害，便會選擇隱藏自己意見，保持沉默，最後支持主流意見的聲音會愈來愈大，而弱勢意見的聲音逐漸消失。《和主流意見不同的人，為什麼會選擇沉默？》，https://www.managertoday.com.tw/glossary/view/55?。

8 "Down 39% in 2022, Meta Platforms Is a Screaming Buy Right Now", https://www.fool.com/investing/2022/02/28/down-39-in-2022-meta-platforms-is-a-screaming-buy.

9 台灣事實查核中心，https://tfc-taiwan.org.tw/；Cofacts 真的假的，https://cofacts.tw/；MyGoPen（麥擱騙），https://www.mygopen.com/。

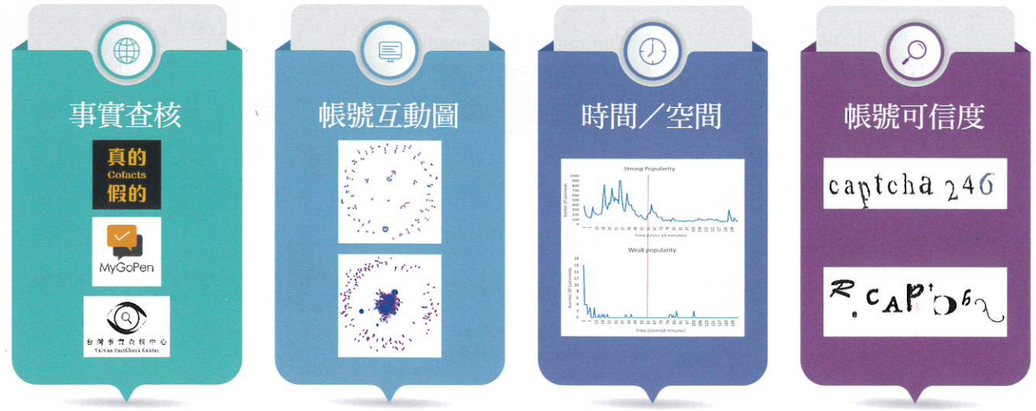


圖 2　目前檢驗訊息可信度的方式

缺點則為平臺掌握絕對權力，且容易侵犯用戶隱私。

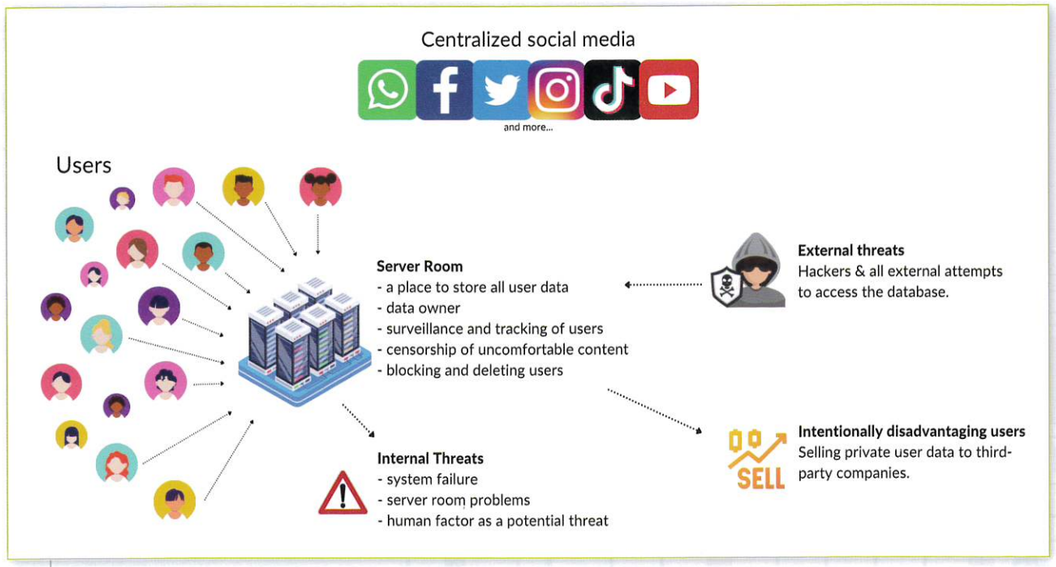
除了帳號偵測與人工審核之外，另可針對可疑訊息的流傳範圍、特徵下手，給定一則訊息，觀察其流傳的頻道（粉絲專頁）、時間、地理位置，以及其參與者彼此間是否有不正常之協同行為。**透過近年發展迅速的人工智慧以及巨量資料等技術**，搜集大量標記的可疑訊息之特徵，運用文字向量化、圖卷積網路等深度學習技術訓練一分類器，**使得電腦自動分類未來訊息是否為可疑新聞**，或者來自地理位置變換快速的發文者，進而自動化調整社群推薦系統權重，在降低可疑文章推送的同時，也同步標記可疑的粉絲專頁以及個人帳號，使得攻擊者做逆向工程以及發動殭屍帳號群的**成本大增**。缺點為只要知道平臺檢測的人工智慧演算法，攻擊者還是有辦法創造出可以混淆分類器的可疑新聞，繞過電腦的檢測，可謂是「道高一尺、魔高一丈」。

傳統中心化社群媒體

為什麼社群媒體平臺安全問題近年來引起關注，我們需要從其特性講起。中心化社群媒體主要由少數人控制和單一集中伺服器運作，而每個社群媒體都有各自不同的訊息審查標準及權限規則，也因如此，使用者在分享或發布訊息時，並不一定能夠暢所欲言，甚至公眾人物可以發布不實訊息去影響群眾。

中心化社群媒體主要有以下特性：伺服器由單一組織掌控 [10]、少數人對於社群

10 伺服器屬於某社群媒體公司，社群媒體的一切運作皆由該伺服器執行，所需資訊均在伺服器統一管理。當有需求者向伺服器尋問時，將可以快速取得所需資訊，然若使用者數量增加到一定程度時，伺服器將面臨擴充性的問題，或是伺服器故障時無法分散風險系統將會整體癱瘓。



中心化社群媒體的伺服器由單一組織掌控，擁有極大的控制權、主觀的審查標準，且資料為封閉、集中式的儲存，恐面臨系統整體癱瘓、使用者資料被出售、言論爭議產生與遭駭客攻擊資料外洩等各項風險。（Source: floyx, https://www.floyx.com/learn-more#section3）

媒體擁有極大的控制權[11]、主觀審查標準[12]與資料為封閉且集中式儲存[13]等，恐將面臨系統整體癱瘓、使用者資料被出售、言論爭議產生與駭客攻擊時資料全數外洩等各項風險。

區塊鏈與社群媒體

區塊鏈技術透過密碼學數位簽章、雜湊函數以及共識獎勵機制來達成四大特性，包括：去中心化、匿名性、不可篡改、共識與獎勵機制。而上述特性恰為當前中心化社群媒體所需要革新的方向。

關於去中心化特性，因為區塊鏈會在全世界擁有多個複本，因此應用於社群平臺發表言論可說是「覆水難收」。[14]在帳號管控方面，區塊鏈社群平臺與比特幣、以太坊系統一樣，只要用戶產生出公私鑰配對，即可以加入討論區且保持匿名性。最後也是最重要的是，**區塊鏈社群平臺改寫了由傳統社群媒體公司壟斷的分潤機制與推薦機制**。高品質的文章由所有用戶決定，用戶不再是一人一票，而是依據名望聲譽（Reputation）與對平臺貢獻高低來決定比例，並且所有程式碼以及獎勵機制全部公開，平臺不會一夕之間改動分潤以及廣告推送模式，**平臺收益可以更準確地回饋至實際對平臺有貢獻的使用者身上**，形成良性循環。

關於區塊鏈社群平臺與傳統社群平臺比較如下表 1 所示。

區塊鏈顛覆訊息推送機制

如圖 1 所示，關於社群媒體的資訊安全問題，最重要也是最常被忽略的是位於最上層的廣告以及推薦系統（Recommendation Systems），因為中心化社群媒體通常為公開上市公司，必須為股東負責；可疑訊息的聳動、談論性高等特性較易引起使用者傳播，也因此時常占據使用者裝置的「熱區」，追根究底，**乃是其複雜私密又龐大的推送引擎機制，才使得可疑新聞的影響愈來愈無遠弗屆**。也因此近年來關於公開社群媒體巨獸公司推送演算法的議題引起學業界廣大的辯論，連全球首富 Elon Musk 都默默收購 Twitter 股票，成為 Twitter 最大股東，意在公開其推薦系統公開原始碼。

相比中心化社群媒體平臺，區塊鏈社群平臺的特色為「使用者即股東」，以全球最大的 Steemit 平臺為例，[15] 盈餘 90% 分配給平臺股東，剩餘的 10% 進入貢獻池，比例分配如下：

- 發文章、評論、回覆等創作者獎勵占 75%
- 給文章投票是否值得推薦，獎勵 15%
- 打包上鏈見證人 10%

基本上，區塊鏈社群平臺篩選內容好壞是由全體用戶投票決定，一則訊息發布

表 1　中心化與區塊鏈社群媒體之比較

特　　性	傳統中心化社群媒體	區塊鏈社群媒體
權威機構認證	由中心化機構背書	全體使用者參與
資料儲存	集中式伺服器	分散各節點儲存
言論審查	機構主觀審查	客觀表決
創作者獎勵機制	封閉、隨時可改動	開放原始碼
隱私性	低	高
實際產品	Twitter、Facebook	Steemit、Matters

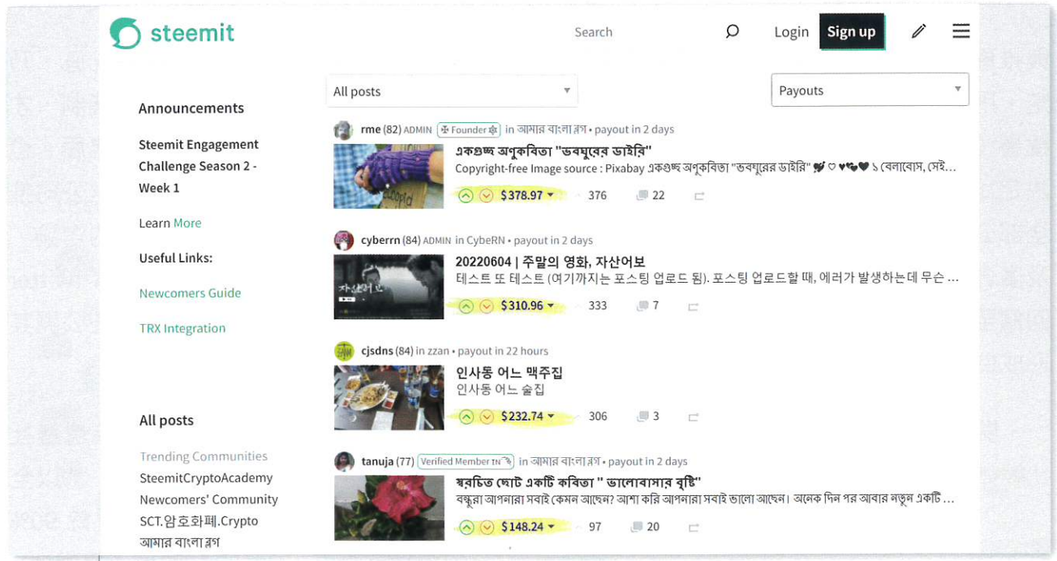
[11] 在社群媒體上所有的使用者資料都掌控在該社群媒體公司，而社群媒體公司擁有刪除使用者資料的權力，甚至將使用者資料出售給利益團體。

[12] 每個中心化的社群媒體對於言論尺寸拿捏的規則標準不一，皆是由該中心化社群媒體主觀單方面認定為不當言論，而當言論的適切性皆由主觀認定時，最後將導致控制使用者言論的爭議產生。

[13] 集中式儲存雖然可使需求者快速獲取資料，但資料無法分散資料流出風險，例如發生駭客攻擊事件將會有很大的可能性資料全數外洩。

[14] 包括 Twitter、Facebook、Instagram 等平臺只要結束運營，或者可由使用者、平臺端刪除某言論。然而在以後的區塊鏈社群平臺所有發言一旦上鏈後，所有的文章、留言皆永久保存。

[15] Steemit 平臺主要運行於 Steem 鏈上，資產共分有 Steem（類似現金）、Steem Power（股權）以及 Steem Dollars（債卷）三類，三種貨幣形式可以透過某些活動互相轉換。



區塊鏈社群平臺的特色為「使用者即股東」，全球最大的 Steemit 平臺盈餘有 90% 分配給股東，其中包含發布訊息的獎勵；而篩選訊息內容好壞則由全體用戶投票決定，訊息品質愈高的人獲得獎勵愈多。（Source: Steemit, https://steemit.com）

出來，讓用戶投票，誰的股權多，占比就愈高，最後訊息品質愈高的人獲得獎勵愈多。為了防止用戶惡意炒作，其引入類似否決票（即有正義用戶舉報）、投票速度限制以及延遲文章獎勵機制，類似各國申請信用卡所需審核的「社會安全碼機制」，任何不良記錄都會永久保存，信用值較低的帳號起不了任何炒作的作用。

總體來說，區塊鏈社群平臺機制相當複雜，且變動極其快速，但其**擺脫了中心化社群媒體固有的封閉框架**，然而是否真為一可靠資訊傳播平臺，仍有待時間以及市場來證明。

社團法人台灣 E 化資安分析管理協會（ESAM）

區塊鏈社群平臺技術，防範假訊息流竄新利器

各種通訊軟體愈來愈多元，接收各種資訊的頻率越來越高，這些平臺默默蒐集使用者資料並且導致許多安全與隱私問題。在治標方面，包括人為識別可疑訊息、定期辦理澄清工作坊、殭屍帳號偵測與移除、運用人工智慧篩選可疑訊息異常傳播時空路徑、訴諸法律管理等措施，皆為有效杜絕有心人士藉社群平臺詐騙、認知作戰，進而瓦解人民、特定組織與國家安全的手段。在治本方面，新興區塊鏈社群平臺技術或可帶來革命性的創新，惟尚須時間證明。