Annex 1: social media and online technological landscape

Social media and online platforms that young people use have various technological factors influencing their social media and online experience. Table 20 outlines some of the key factors which are shaping social media and the online environment, and the potential benefits and risks of this technology.

Table 20: Social media and online technological landscape.

Technology	Summary	Benefits	Risks
1.1 Curation algorithms.	Social media companies use algorithms that curate what content a user sees. Algorithms are used to ensure that the content is relevant and of interest to the user,[52,451] and that they stay on the platform for a longer period of time. Curation algorithms have large influence on Facebook feeds,[451] TikTok's For You page,[52] and Instagram's explore and feed.[452,453] They are also used by search engines to help dictate the search results which appear.	• Individuals see content which is of interest and more relevant to them.[52,454] • People can find their community online, with evidence that minorities, the LGBTQI+ community and women can find content and safe spaces online.[455]	• Political content that elicits strong responses, including negative, is more likely to be rewarded by the algorithm.[454] Divisive content which captures attention is therefore more likely to be spread on platforms which can further entrench ideas.[456] • Can create individualised curated content online, meaning that people see content dependent on their ideas and interactions online, which can vary drastically between individuals, as acknowledged by some platforms,[52,455,456] fueling polarisation. There is some debate as to the extent this occurs[457] and some research questions the validity of the so called filter bubble hypothesis.[82]
1.2 Moderation.	Moderation refers to removing or managing content or people who breach user and community guides of a social media platform. Content which breaches user guidelines includes that which incites violence, contains nudity, hate speech, or polluted information. Moderation is split into two types, ex-post and algorithmic, with most social media companies using both.[85,458-460]	• Content can be removed which breaches community guidelines.[458,459] • Removes extremist and terrorist content.[466]	• Some groups feel they are being disproportionately moderated.[467] • Disagreement about what and who should be moderated.[468] • Currently, processes are largely dictated by social media companies, with few democratic states regulating this process.[458]

	Exact techniques of moderation are unknown, due to social media companies largely being closed-book. Some moderation techniques include: removing posts, removing accounts, link to official information, blurring graphic content, content warnings, and removing hashtags.[459,461-464] Advertisements on social media platforms can be moderated.[465]		• Extremists who are moderated can move to other platforms which aren't moderated, with some evidence they become more toxic on these platforms.[464] • Posts with warning tags have the potential unintentional consequence of being ignored, or make the user believe the content more, or increase the desirability of the videos for young people.[469] • Users can circumvent hashtag bans by adapting the hashtags and descriptions they use.[461]
	Ex-post moderation. Moderation which occurs after content has been posted, relies on people reporting the content; this essentially crowdsources moderation, with human review of the reported content. In some cases AI is also doing this.[460]	• Allows for human understanding of nuance or cultural references. • Involves wider online users in the process, ensuring that members of the community contribute to moderation.	• Time constraints on moderators mean they have only tens of seconds to moderate a post.[460] • The individual moderator has a level of subjectivity, which impacts on moderation practices.[460] • Relies on users reporting content.[462]
	Algorithmic and AI moderation. Moderation occurring through algorithms, largely occurring as content is posted, has a broad intent for algorithms to uphold community guidelines.[462] AI is also increasingly being used, especially on platforms with large scales.[468]	• Intercepts content before being posted, meaning that it doesn't reach other users.[470] • The technology will continue to develop and strengthen over time, ensuring greater accuracy into the future. • Allows for much larger scale than ex-post moderation.[462]	• Moderation is only as good as the algorithm and the guidelines it seeks to uphold.[462] • Algorithms are likely to make hundreds and possibly thousands of mistakes a day.[460] • Can have unintended impacts on content which doesn't breach community guidelines, e.g., changes in Tumblr's algorithm meant sex education material was removed, not just explicit images.[460,467] • Algorithms can miss or misinterpret slang or country specific language.[470]

			• Currently non-transparent and difficult to audit or regulate.[470] • Limited transparency can fuel lack of trust in the moderation process.[85]
1.3 Deepfakes.	Deepfakes refer to the ability through AI and technology to alter, superimpose, or change video, images and audio, usually changing those who appear in the content.[471] The AI technology is rapidly evolving, meaning convincing deepfakes can now be made easily and cheaply.[471]	• Innocent use of deepfakes for creative content.	• Risk of being used as polluted information to sow political divisions.[455] • Can increase the effectiveness of cyber enabled information warfare, being utilised by foreign actors to interfere in domestic democracy.[472] • Increasingly convincing, making it difficult for an average member of the public to identify media as a deepfake.[473] • Individuals', including celebrities', images are being transplanted onto pornographic materials, with some being circulated online. This constitutes a form of image based abuse and online harassment.[474]
1.4 Community and user guidelines.	Community and user guidelines are produced by social media companies, and set out expected behaviour on their platform. They are used in moderation, with breaches of the guidelines grounds for removal of content or a user.[459] Community and user guidelines are publicly available and generally are extensive[459,475-477] but how they are implemented is more opaque.	• Gives a set of rules that users are expected to abide by and a mandate to remove content which isn't in line with this.[477]	• Currently dictated by social media companies.[478] • Not uniform, each social media platform has slightly different guidelines.[478] • The extent and manner in which they are enforced has remained largely invisible.[470]
1.5 Paid advertising.	Social media companies employ paid advertising as the key revenue stream.[479] By collecting users data, they can ensure that paid advertisers can target groups who are most likely to engage with an advertiser or buy a product.[479-482]	• Advertisers, including social and community groups can reach communities.[480] • Can help with public health or emergency messaging.[483]	• Issues and concern of data privacy.[484] • Advertisers as the key source of revenue are the main customer for social media companies, not users. Social media companies are focused on maintaining and expanding

		• Social media users have ads which are more relevant to them.[479,480,482]	paid advertisers with the users the secondary focus.[479] • Can include polluted information (in spite of advertising moderation) particularly when there is complexity – including greenwashing or unverified medicinal claims.[485-487] • Has been used by foreign countries to influence domestic elections.[485] • Users may struggle to identify what is paid advertising and what isn't.[488]
1.6 ChatGPT and related LLM.	AI is increasingly being used both online generally and on social media platforms. ChatGPT is an AI Chatbot released by OpenAI in November 2022 which has made waves in both the academic community and general public.[313] ChatGPT creates realistic sounding text when responding to prompts, it creates the text using neural networks which digest huge amounts of existing human-generated text. Each output is unique, meaning it isn't picked up by plagiarism checkers.[489] The outputs of ChatGPT are imperfect and there are still substantive gaps between quality writing and ChatGPT outputs,[490] but this gap is likely to continue to close as the AI develops and expands.[489]	• LLMs can access information quickly drawing information together from multiple sources.[489] • Produces original content each time.[313] • Can be guided by the user to produce more specific or accurate outputs.[489]	• Concern about plagiarism and using LLMs for work, in the education system, and academic community.[489,490] • The very real near-term potential that AI/ChatGPT will be able to write as well or better than humans with the potential for AI to takeover areas of research.[313] • LLMs cannot currently adequately distinguish between false and accurate information.

References

52. TikTok. How TikTok recommends videos #ForYou. (2020). https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you.

82. Ross Arguedas, A., Robertson, C.T., Fletcher, R. & Nielsen, R.K. (2022), Echo chambers, filter bubbles, and polarisation: A literature review The Royal Society, University of Oxford, and Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-01/Echo_Chambers_Filter_Bubbles_and_Polarisation_A_Literature_Review.pdf

85. Ozanne, M., Bhandari, A., Bazarova, N.N. & DiFranzo, D. (2022), Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, 9(2): 1-13. https://doi.org/10.1177/20539517221115666

313. Gordijn, B. & Have, H.T. (2023), ChatGPT: Evolution or revolution? *Medical, Health Care and Philosophy*. 26: 1-2. https://doi.org/10.1007/s11019-023-10136-0

451. Facebook, How feed works. Retrieved 21 November 2022 from https://www.facebook.com/help/1155510281178725

452. Instagram, How Instagram Feed works. Retrieved 21 November 2022 from https://help.instagram.com/1986234648360433

453. Instagram, How posts are chosen for Explore on Instagram. Retrieved 21 November 2022 from https://help.instagram.com/487224561296752

454. Cho, J., Ahmed, S., Hilbert, M., Liu, B. & Luu, J. (2020), Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media*, 64(2): 150-172. https://doi.org/10.1080/08838151.2020.1757365

455. Smith, B. & Browne, C.A. *Social media: The freedom that drives us apart.* In Tools and Weapons: The promise and peril of the digtial age, Hodder and Stoughton: London, 2019.

456. Van Bavel, J.J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. (2021), How social media shapes polarization. *Trends in cognitive science*, 25(11): 913-916. https://doi.org/10.1016/j.tics.2021.07.013

457. Jones-Jang, S.M. & Chung, M. (2022), Can we blame social media for polarization? Counter-evidence against filter bubble claims during the COVID-19 pandemic. *New Media & Society*: 1-20. https://doi.org/10.1177/14614448221099591

458. Myers West, S. (2018), Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11): 4366-4383. https://doi.org/10.1177/1461444818773059

459. TikTok, Community guidelines TikTok. Retrieved 22 November 2022 from https://www.tiktok.com/community-guidelines?lang=en#29

460. Young, G.K. (2021), How much is too much: The difficulties of social media content moderation. *Information & Communications Technology Law*, 31(1): 1-16. https://doi.org/10.1080/13600834.2021.1905593

461. Gerrard, Y. (2018), Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12): 4492-4511. https://doi.org/10.1177/1461444818776611

462. Thach, H., Mayworm, S., Delmonaco, D. & Haimson, O. (2022), (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*: 1-20. https://doi.org/10.1177/14614448221109804

463. Morrow, G., Swire-Thompson, B., Polny, J.M., Kopec, M. & Wihbey, J.P. (2022), The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10): 1365-1386. https://doi.org/10.1002/asi.24637

464. Jhaver, S., Boylston, C., Yang, D. & Bruckman, A. (2021), Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1-30. https://doi.org/10.1145/3479525

465. Liu, Y., Pinar Yildirim, T. & Zhange, J. (2021), Implications of revenue models and technology for content moderation strategies. *Social Science Research Network*, 41(4): 663-869. https://doi.org/10.2139/ssrn.3969938

466. Ganesh, B. & Bright, J. (2020), Countering extremists on social media: Challenges for strategic communication and content moderation. *Policy & Internet*, 12(1): 6-19. https://doi.org/10.1002/poi3.236

467. Haimson, O.L., Delmonaco, D., Nie, P. & Wegner, A. (2021), Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5: 1-35. https://doi.org/10.1145/3479610

468. Gillespie, T. (2020), Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2): 1-5. https://doi.org/10.1177/2053951720943234

469. Sharevski, F., Devine, A., Jachim, P. & Pieroni, E. Meaningful context, a red flag, or both? Preferences for enhanced misinformation warnings among US Twitter users. Presented at: 2022 European Symposium on Usable Security, (2022).

470. Gorwa, R., Binns, R. & Katzenbach, C. (2020), Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 1-15. https://doi.org/10.1177/2053951719897945

471. Liv, N. & Greenbaum, D. (2020), Deep fakes and memory malleability: False memories in the service of fake news. *American Journal of Bioethics Neuroscience*, 11(2): 96-104. https://doi.org/10.1080/21507740.2020.1740351

472. Paterson, T. & Hanley, L. (2020), Political warfare in the digital age: Cyber subversion, information operations and 'deep fakes'. *Australian Journal of International Affairs*, 74(4): 439-454. https://doi.org/10.1080/10357718.2020.1734772

473. Basch, C.H., Meleo-Erwin, Z., Fera, J., Jaime, C. & Basch, C.E. (2021), A global pandemic in the time of viral memes: COVID-19 vaccine misinformation and disinformation on TikTok. *Human Vaccines and Immunotherapeutics*, 17(8): 2373-2377. https://doi.org/10.1080/21645515.2021.1894896

474. Maddocks, S. (2020), 'A deepfake porn plot intended to silence me': Exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 7(4): 415-423. https://doi.org/10.1080/23268743.2020.1757499

475. Meta, Facebook community standards. Retrieved 21 November 2022 from https://transparency.fb.com/en-gb/policies/community-standards/

476. Twitter, The Twitter rules. Retrieved 21 November 2022 from https://help.twitter.com/en/rules-and-policies/twitter-rules

477. Instagram, Community guidelines. Retrieved 21 November 2022 from https://help.instagram.com/477434105621119

478. Jiang, J.A., Middler, S., Brubaker, J.R. & Fiesler, C. (2020), Characterizing community guidelines on social media platforms. *Conference companion publication of the 2020 on computer supported cooperative work and social computing*: 287-291. https://doi.org/10.1145/3406865.3418312

479. Instagram, Instagram terms and imprint. Retrieved 22 November 2022 from https://help.instagram.com/581066165581870/

480. Twitter, Twitter ads targeting. Retrieved 22 November 2022 from https://business.twitter.com/en/advertising/targeting.html#audience-types

481. TikTok, Bring your brand's voice to life with our solutions. Retrieved 22 November 2022 from https://www.tiktok.com/business/en/solutions

482. Facebook, Target future customers and fans. Retrieved 22 November 2022 from https://en-gb.facebook.com/business/ads

483. Wolkin, A.F., Schnall, A.H., Nakata, N.K. & Ellis, E.M. (2019), Getting the message out: Social media and word-of-mouth as effective communication methods during emergencies. *Prehospital and Disaster Medicine*, 34(1): 89-94. https://doi.org/10.1017/S1049023X1800119X

484. Jung, A.R. (2017), The influence of perceived ad relevance on social media advertising: An empirical examination of a mediating role of privacy concern. *Computers in Human Behavior*, 70: 303-309. https://doi.org/10.1016/j.chb.2017.01.008

485. Kim, Y.M., Hsu, J., Neiman, D., Kou, C., Bankston, L., Kim, S.Y., Heinrich, R., Baragwanath, R. & Raskutti, G. (2018), The stealth media? Groups and targets behind divisive issue campaigns on Facebook. *Political Communication*, 35(4): 515-541. https://doi.org/10.1080/10584609.2018.1476425

486. Naderer, B. & Opree, S.J. (2021), Increasing advertising literacy to unveil disinformation in green advertising. *Environmental Communication*, 15(7): 923-936. https://doi.org/10.1080/17524032.2021.1919171

487. Goodwin, I. (2022), Programmatic alcohol advertising, social media and public health: Algorithms, automated challenges to regulation, and the failure of public oversight. *International Journal of Drug Policy*, 109: 103826. https://doi.org/10.1016/j.drugpo.2022.103826

488. Parasnis, E. (2022), The implications of social media for adolescent critical thinking from an information and advertising literacy context: A brief review. *The Serials Librarian*, 83(1): 9-15. https://doi.org/10.1080/0361526x.2022.2030850

489. Else, H. (2023), Abstracts written by ChatGPT fool scientists. *Nature*, 613: 423. https://doi.org/10.1101/2022.12.23.521610

490. Thorp, H. (2023), ChatGPT is fun, but not an author. *Science*, 379(6630): 313. https://doi.org/10.1126/science.adg7879