31 (2) 2023 IIUMLJ 155 - 178

INTERMEDIARY'S LIABILITY: TOWARDS A SUSTAINABLE ARTIFICIAL INTELLIGENCE-BASED CONTENT MODERATION IN MALAYSIA

Mahyuddin Daud*

Ida Madieha Abd Ghani Azmi**

ABSTRACT

Intermediaries enjoy a 'safe harbour' from civil or criminal liability should they host illegal and harmful third-party online content, subject to requirements from national laws. The line between immunity and liability becomes hazy when intermediaries appear to assume the role of content creators, hence risk being characterised as publishers. The paper analyses whether such liability may be diminished if intermediaries adopt an artificial intelligence-based content moderation system. Through comparative case analyses of Mkini, Delfi, Bunt and Godfrey, the research questioned the relevance of the legal defence granted in the Communications and Multimedia Act 1998 and the Content Code. The article analysed the Federal Court's finding of liability for Mkini and asked whether it signals the right way forward that may ignore the fundamental right to express opinion of public interest. Despite the advances in artificial intelligence-based content moderation, it remains to be seen if algorithms can easily contain illegal and harmful content.

Keywords: Artificial Intelligence, Content Moderation, Intermediary Liability, Content Regulation, Safe Harbour.

* Department of Civil Law, Ahmad Ibrahim Kulliyyah of Laws, International Islamic University Malaysia. Email: mahyuddin@iium.edu.my.

** Department of Civil Law, Ahmad Ibrahim Kulliyyah of Laws, International Islamic University Malaysia. Email: imadieha@iium.edu.my.

LIABILITI PENGANTARA: KE ARAH MODERASI KANDUNGAN BERASASKAN KECERDASAN BUATAN YANG LESTARI DI MALAYSIA

ABSTRAK

Pengantara menikmati `pelabuhan selamat' daripada liabiliti sivil atau jenayah sekiranya mereka mengehoskan kandungan dalam talian pihak ketiga yang menyalahi undang-undang dan berbahaya, tertakluk pada keperluan undang-undang negara. Garis antara imuniti dan liabiliti menjadi kabur apabila pengantara kelihatan memainkan peranan sebagai pencipta kandungan, justeru berisiko disifatkan sebagai penerbit. Makalah ini menganalisis sama ada liabiliti sedemikian boleh dikurangkan jika pengantara mengamalkan sistem moderasi kandungan berasaskan kecerdasan buatan. Melalui analisis kes perbandingan Mkini, Delfi, Bunt dan Godfrey, kajian mempersoalkan kaitan perlindungan undang-undang yang diberikan dalam Akta Komunikasi dan Multimedia 1998 dan Kod Kandungan. Artikel ini menganalisis penemuan Mahkamah Persekutuan untuk meletakkan liabiliti terhadap Mkini dan bertanya sama ada ia memberi isyarat yang betul yang mungkin mengabaikan hak untuk menyatakan pendapat berunsurkan kepentingan awam. Walaupun terdapat kemajuan dalam moderasi kandungan berasaskan kecerdasan buatan, masih perlu dilihat sama ada algoritma boleh menyekat kandungan yang menyalahi undang-undang dan berbahaya dengan sewajarnya.

Kata kunci: Kecerdasan Buatan, Moderasi Kandungan, Liabiliti Pengantara, Peraturan Kandungan, Pelabuhan Selamat.

INTRODUCTION

Automated decision-making and predictive analytics through artificial intelligence (AI), along with rapid advances in sensor technology and robotics may revolutionise how individuals, communities, governments, and private entities perceive and respond to technological change. Sustainability requires change, and businesses are learning how to consolidate new technology and new approaches to advance their social-ecological framework and development. Sustainability is being shaped by technology, which enables enhanced levels of productivity and efficiency. Despite increased interest and

deployment of AI technologies in sustainability-critical domains, few have explored systemic hazards it may cause.[1]

Intermediary liability is an area with uncertain development and inconsistent standards across the globe. With each country having a distinct threshold for intermediary's liability as opposed to publishers, many of them had inadvertently shouldered the responsibility for hosting illegal third-party content online.[2] The global standard is that an intermediary enjoys a certain level of immunity against third party publications[3] as they are considered to be just an 'intermediary' as the title itself suggests. However, with active encouragement of posting user-generated elements to increase web interactivity, the legal position of content creators, whether they are still deemed to be passive conduits or active publishers remains far from clear.

Recent works of the Organisation for Economic Co-operation and Development and article 19 categorises intermediaries as organisations who "bring together or facilitate transactions between third parties on the Internet.[4] These organisations provide access, host, transmit and index content, products and services originated by third parties on the Internet or provide Internet-based services to third parties. The following are categories of intermediaries as illustrated by the OECD:

1. Internet access and service providers (ISPs);
2. Data processing and web hosting providers, including domain name registrars;
3. Internet search engines and portals;

---

[1] Victor Galaz et al., "Artificial Intelligence, Systemic Risks, and Sustainability," *Technology in Society* 67 (November 1, 2021): 101741, https://doi.org/10.1016/J.TECHSOC.2021.101741., accessed December 1, 2022.

[2] Suzi Fadhilah Ismail, Ida Madieha Abdul Ghani Azmi, and Mahyuddin Daud, "Transplanting the United States' Style of Safe Harbour Provisions on Internet Service Providers Via Multilateral Agreements: Can One Size Fit All?," *IIUM Law Journal* 26, no. 2 (2018), https://journals.iium.edu.my/iiumlj/index.php/iiumlj/article/view/396., accessed December 1, 2022.

[3] See, for example, Articles 12-15 of the EU E-Commerce Directive 2000.

[4] The Organisation for Economic Co-operation and Development, "The Economic and Social Role of Internet Intermediaries," 2010, accessed December 1, 2022, www.oecd.org/dataoecd/49/4/44949023.pdf.

4. E-commerce intermediaries, where these platforms do not take title to the goods being sold;
5. Internet payment systems; and
6. Participative networking platforms, which include Internet publishing and broadcasting platforms that do not themselves create or own the content being published or broadcast.

Internet intermediaries are groups of organisations including Internet service providers, Web hosting providers, social media platforms and search engines, but excludes content providers. Among those listed in the group are (i) Internet service providers; (ii) Web hosting providers; (iii) social media platforms and (iv) search engines.[5] The standard position is that Internet intermediaries are considered the mere 'middleman' and they do not actively create content for netizens. This is also similar to the WhatsApp service whereby it is merely a social media platform that enables netizens to connect. The instantaneous chat platform does not provide any content whatsoever.

It is common for content creators to provide a certain section on their page or article that invites visitors to comment. For online news portals, they report news articles, and provide comments section at the bottom of the said article. Social media platforms have adopted this trend whereby massive virtual spaces are provided for users to input their thoughts through comments, and to add content of their own. Users may upload pictures, videos, music, on top of still-texts. Social media platforms do not usually pre-review what the users contribute to their pages due to the massive amount of user-generated content they receive per second, hence will act upon complaints.[6] The interactions between the platform and the users create bondage between them and constitute the main attraction for users to continue engaging with the platform. These interactions become crucial in retaining customer loyalty as users feel welcomed to air their views. As these interactions

[5] ARTICLE19, "Internet Intermediaries: Dilemma of Liability Q and A," ARTICLE 19, 2013, accessed December 1, 2022, https://www.article19.org/resources.php/resource/37243/en/internet-intermediaries:-dilemma-of-liability-q-and-a.

[6] Facebook, "Statement of Rights and Responsibilities," Facebook, 2013, https://www.facebook.com/legal/terms; YouTube, "Community Guidelines Strikes," YouTube, 2018, accessed December 1, 2022, https://support.google.com/youtube/answer/2802032?hl=en.

become massive, the platforms felt the need to deploy some form of content moderation based on certain algorithms that determine how contents were to appear on their website and basically can remove certain content if deemed necessary. It is on the basis of the dynamism of online traffic that 'intermediaries' qualify for the safe harbour for any third-party materials posted on their website.

Since the OECD and Article 19 simply set out to classify 'who' and 'which organisation qualifies as intermediaries, the liability that these intermediaries face in the hosting of content differs from one country to another, depending on their legislative provision, and domestic court's interpretation. In Malaysia, the position of intermediaries is set under the Content Code on the basis of three types of liabilities, i.e., i) mere conduit ii) caching and iii) hosting. Hosting is the one that subjects the intermediaries to a multitude of liabilities including the possible liability as publisher and editor. As the liability for hosting varies according to the type of content, the moot question remains, what if the intermediary deploys a smart system to moderate content. This article continues the discussion on the use of AI as a content moderator.

AI as content moderator or web filter

Since intermediaries who assume the role of content creators may risk being labelled as publishers, the ensuing issue is whether the use of artificial intelligence may be the way forward to reduce legal risks? Is it possible to employ machine learning to filter content so as to minimise risk for content creators, especially for a third-party content that was posted on their websites?

In this regard, Job Turner defined AI workability as the capacity of a machine or computer programme to behave intelligently in the same manner as a human being would.[7] As a result, human intellect becomes a proxy for what AI accomplishes. Intelligence is the capacity to reason abstractly, logically, and consistently, to discover, lay, and see-through correlations, to solve problems, to discover rules in apparently disordered material, to solve new tasks, to adapt flexibly to new situations, and to learn independently, without the need for explicit

[7] Job Turner, *Robot Rules: Regulating Artificial Intelligence* (London: Palgrave Macmillan, 2019), 7–8.

and comprehensive instruction.[8] The successful model of AI integration has been seen in social media marketing.

Simultaneously, AI is becoming more integrated into many facets of social media. Although AI will never completely replace the human touch, it is improving the amount and quality of online interactions between businesses and their consumers. Machine learning, a subset of AI, may be used in the following four ways by businesses to generate efficient social media marketing strategies[9] – that may be used as an effective mechanism for content moderation and social media monitoring. Twitter and Instagram have included built-in analytics tools that may track the number of likes, comments, link clicks, and video views for previous postings. Similar social media analysis and management services may be provided by third-party providers, which can teach organisations about their consumers, including demographic information and the best times to publish. Because social media algorithms favour more recent posts over older posts, businesses can use this information to strategically schedule their posts during peak times or a few minutes before peak times.[10]

An AI that is capable of assessing the sentiment of text data, is a process known as sentiment analysis.[11] To connect social media data with predetermined sentiment categories such as positive, negative, or neutral, the procedure employs both natural language processing (NLP) and machine learning.[12] The system may then train agents to recognise the underlying feelings in fresh texts. Sentiment analysis may be used in social media and customer service to get input on a new product, service, or design. Businesses may use sentiment analysis to see how

---

[8] Turner, *Robot Rules: Regulating Artificial Intelligence*.

[9] Telus International, "What Are Search And Recommendation Systems In ML?," Telus International, February 1, 2021, https://www.telusinternational.com/articles/4-ways-machine-learning-can-enhance-social-media-marketing.

[10] Telus International, "What Are Search And Recommendation Systems In ML?," Telus International, accessed February 1, 2021, https://www.telusinternational.com/articles/4-ways-machine-learning-can-enhance-social-media-marketing.

[11] Turner, *Robot Rules: Regulating Artificial Intelligence*, 168; Telus International, "What Are Search And Recommendation Systems In ML?"

[12] Telus International, "What Are Search And Recommendation Systems In ML?"

people feel about their competition or hot subjects in their sector. Using sentiment analysis, content creators may be able to analyse in advance, which posts may invite illegal content. Hence AI may be used as an early warning system and implement automated takedowns once a list of objectionable words has been detected in the comments section.

For social media marketing, image recognition is essential. Without any accompanying text, image recognition employs machine learning to train computers to recognise a brand logo or photographs of certain items. When users share images of a product without immediately stating the brand or product name in a caption, this might be valuable for businesses. Customers may also post a snapshot of your product with the phrase "Where can I get this?" on social media. When companies see this, they may utilise it to offer targeted marketing to that individual, or simply comment on the post to provide an explanation, resulting in higher customer satisfaction. Such machine learning may also be used to detect if third party users post inappropriate pictures in the comments section that can attract unnecessary liability for content creators.

On the other hand, 'chatbots' are AI applications that replicate real-life discussions. Chatbots can be incorporated into websites like online retailers, or they can be accessed via a third-party chat network like Facebook Messenger, Twitter, or Instagram's direct messaging. Chatbots are more likely to boost customer satisfaction for organisations with a mostly youthful consumer base. More than 60% of millennials have used chatbots, and 70% of them have had pleasant encounters with them.[13] Chatbots may be used in a variety of scenarios, not only when a consumer has a specific inquiry or complaint. Estee Lauder has a chatbot incorporated in Facebook Messenger that utilises facial recognition to help consumers choose the proper foundation shade[14], while Airbnb has utilised Amazon Alexa to greet visitors and direct them to nearby sights and restaurants.[15] The same goes for

[13] Telus International.

[14] Robert Williams, "Estée Lauder's AR Chatbot Offers Advice on Lipstick Colors | Marketing Dive," Marketing Dive, July 14, 2017, https://www.marketingdive.com/news/estee-lauders-ar-chatbot-offers-advice-on-lipstick-colors/447096/.

[15] Eddie Star, "The Game-Changing Role of AI in Airbnb's Success," Linkedin, 2023, https://www.linkedin.com/pulse/game-changing-role-ai-airbnbs-success-eddie-starr/.

Airasia[16] which employs a chatbot to address customer inquiries. Perhaps in future, chatbots can be used to filter the third-party content as a measure to protect content creators against liability for third-party publication.[17]

What are Machine Learning and Deep Learning in AI?

Machine learning is a subfield of artificial intelligence and a technique for developing systems that solve problems. Prior to the advent of machine learning, programmers manually coded instructions for obtaining a desired output by utilising a certain input. Statistical techniques enable us to teach computers to learn without requiring a precise set of rules. To accomplish this, we expose our system to a large number of samples - ranging from a few hundred to several million - until it ultimately begins to learn and respond (or predict) more accurately over time. Machine learning systems are notoriously specialised, frequently handling only a single type of problem. This might be through internet advertising bidding, detecting credit card purchase fraud, or even recognising malignant skin cells. Many are now capable of performing these jobs at par with or even better than human professionals, and at far greater scales.

Deep learning is one of several machine learning approaches. This ground-breaking technology is based on sophisticated systems called neural networks that replicate the structure and function of the brain to perform pattern recognition. They are composed of artificial neurons connected to one another. The networks are constructed using many layers of these neurons to generate sophisticated designs that aid the system in capturing and recognising patterns. When a large number of these neuronal layers are loaded, the network becomes 'deep', which is why the phrase 'deep learning' is used.[18]

[16] Capital A Berhad, doing business as 'AirAsia', is a Malaysian low-cost airline with its headquarters in Kuala Lumpur. By fleet size and destination coverage, it is Malaysia's largest airline. AirAsia conducts scheduled domestic and international flights to over 165 cities in 25 countries.

[17] Telus International, "What Are Search And Recommendation Systems In ML?".

[18] Turner, *Robot Rules: Regulating Artificial Intelligence*, 274.

These systems have demonstrated outstanding outcomes with excellent precision and reliability and have thus attracted interest among data scientists in recent years. Deep neural networks are extremely expensive to train, and computers at the time lacked computational power. Additionally, they perform better when given a huge amount of data to train on. That is, data in megabytes or gigabytes. Deep learning has acquired a lot of traction among enterprises and researchers now that computer storage (hard drives and SSDs) is inexpensive and substantially more powerful (both CPUs and Graphics Processing Units). Every individual now can train a basic deep neural network with the right computer.[19] Machine learning algorithms are trained on sample postings to identify trends in text or images. They are capable of deciphering minute details and accurately returning the most pertinent answers to questions.[20] Since machine learning relies on examples to discern patterns, it can learn to categorise new posts in any language as long as these articles are correctly tagged with the intended prediction.

With services such as Instagram, Snapchat, and Pinterest, the social web has become increasingly visual. Posts on these platforms are primarily visual, with only a few hints in the text's content. As a result, recognising what was contained in these posts was previously nearly difficult. Fortunately, this is another instance where deep learning comes to the rescue. These algorithms are now capable of recognising logos, faces, and objects in both still and moving images.

How social media platforms use AI

YouTube employs artificial intelligence to promote videos, regulate objectionable content, and assist video creators by providing automated transcription and video effects. YouTube has more than two billion

[19] Gaurav Batra Zach Jacobson Siddarth Madhav Andrea Queirolo Nick Santhanam, "Artificial-Intelligence Hardware: New Opportunities for Semiconductor Companies," Mckinsey, 2018, https://www.mckinsey.com/~/media/McKinsey/Industries/Semiconductors/Our Insights/Artificial intelligence hardware New opportunities for semiconductor companies/Artificial-intelligence-hardware.ashx.

[20] Santhanam.

monthly active users.[21] As a result, humans are unable to moderate the platform, necessitating the use of AI. Artificial intelligence is capable of flagging and removing problematic content in real-time. This enables the automated removal of harmful content before it is seen by a large number of people.

Instagram, like its parent company Facebook, employs artificial intelligence to recommend content and advertisements.[22] As with Facebook, this AI is largely influenced by user preferences and behaviour. Instagram's AI evaluates one's participation on the site and forecasts which piece of content he will engage with the most in the future.[23] Instagram arguably does not appear to have the same issue with harmful content as YouTube. However, it also makes use of artificial intelligence to identify and stop spam bots that are used to farm for likes, clicks, and engagement.[24]

TikTok is a social networking platform that prioritises artificial intelligence. It was designed from the ground up to rely on artificial intelligence. Indeed, AI recommends every piece of content presented on TikTok.[25] Each time one likes or engage with, or detest a piece of content, TikTok's AI improves. TikTok's AI accumulates behavioural data at a breakneck pace thanks to its billion users. The more data that the system possesses, the more accurate its forecasts will be. The more

[21] Rohit Shewale, "YouTube Statistics For 2023 (Demographics & Usage)," Demandsage, 2023, https://www.demandsage.com/youtube-stats/#:~:text=YouTube has more than 2.70,via its Website and Apps.

[22] The Instagram, "Terms of Use," The Instagram, accessed December 1, 2022, http://instagram.com/about/legal/terms/.

[23] Facebook, "How Does Facebook Use Artificial Intelligence to Moderate Content? | Facebook Help Centre," Facebook, September 1, 2023, https://www.facebook.com/help/1584908458516247.

[24] Instagram, "How Instagram Uses Artificial Intelligence to Moderate Content," Instagram, 2023, https://help.instagram.com/423837189385631/?helpref=uf_share.

[25] Net Kohen, "How Your Business Can Take Advantage Of TikTok's Success With AI And ML," Forbes, 2022, https://www.forbes.com/sites/forbestechcouncil/2022/12/09/how-your-business-can-take-advantage-of-tiktoks-success-with-ai-and-ml/?sh=536b50646f93.

accurate its predictions are, the more engaged the user is with TikTok's content.[26]

Notably, automated filtering is not mandated by law, as many laws around the world impose the obligation to take down materials only upon notice by a third party. This hands-off approach augurs well with freedom of expression which is cherished on the Internet. However, the hands-off approach has become more and more not preferred as online interactions boom and is infested with unlawful and inappropriate behaviour.[27] In some cases, intermediaries play a proactive role in content moderation, and in some cases, they play a passive role or even reactive role.[28] The ensuing issue is then if by applying moderation software the intermediaries are turned into publishers and editors and thus not eligible for the 'safe harbour' guaranteed under the law, then more content hosts would choose to not do anything in the face of online unlawfulness and abuse (to retain their immunity).

The Good Samaritan immunity that originated from Section 230 of the US Communication Decency Act should be rolled out to encourage the widespread use of content moderation. Section 230 itself is often referred to as the 'Good Samaritan' clause or immunity. Section 230 of the Communications Decency Act is a crucial piece of legislation in the United States that was enacted in 1996. It provides certain legal protections and immunities to online platforms and internet service providers (ISPs) regarding content posted by users on their platforms. These protections are often summarized as the 'Good Samaritan' clause because they aim to encourage online platforms to moderate and remove offensive or harmful content without fear of legal

[26] Mike Kaput, "What Is Artificial Intelligence for Social Media?," Marketing AI Institute, accessed January 17, 2022, https://www.marketingaiinstitute.com/blog/what-is-artificial-intelligence-for-social-media.

[27] Mahyuddin Daud, *Internet Content Regulation : Contemporary Legal and Regulatory Issues in the Changing Digital Landscape* (Gombak: IIUM Press, 2019).

[28] Lilian Edwards, "The Role and Responsibility of Internet Intermediaries in the Field of Copyright and Related Rights," 2011, http://www.wipo.int/copyright/en/doc/role_and_responsibility_of_the_internet_intermediaries_final.pdf.

repercussions.[29] The European Commission has adopted the same approach in its Communication of September 2017 as the appropriate policy for tackling online content.

The European Parliament, in a report on the use of algorithms for online content filtering, acknowledges that social media dwells heavily on automated filtering to monitor the substantial number of materials that are posted online.[30] The report concedes that most of the systems adopted for content filtering are not error-free. The report recommends that any host that engages in content filtering should not be deprived of whatever immunities or other advantages granted to the providers (Good Samaritan clauses).[31] The report acknowledges that most moderation done by social media (e.g. Google, Facebook or Twitter) is essential to minimise 'bad information', harmful and anti-social behaviour from the online community.[32] These moderations are done for the greater good and not to reduce the free space for expression, but rather to regulate good behaviour online and mitigate harm.[33] In that sense, those platforms act as Good Samaritan and therefore should be totally absolved from liability in the event that the moderation fails to capture contents that are harmful.

The report recommends a specific provision exempting the hosting of user-generated content if they fail to take reasonable measures to prevent unlawful harm to third parties, even with the adoption of algorithm-based content moderation. This is because the system is far from perfect, particularly with regard to toxic speech.[34]

---

[29] Congressional Research Service, "Section 230: An Overview" (United States of America, 2021), chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://crsreports.congress.gov/product/pdf/R/R46751.

[30] Giovanni Sartor and Andrea Loreggia, "The Impact of Algorithms for Online Content Filtering or Moderation," European Parliament Policy Department for Citizens' Rights and Constitutional Affairs Policies, Directorate-General for Internal, 2020, accessed December 1, 2022, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf.

[31] Sartor and Loreggia.

[32] Sartor and Loreggia, "The Impact of Algorithms for Online Content Filtering or Moderation."

[33] Congressional Research Service, "Section 230: An Overview."

[34] Sartor and Loreggia, "The Impact of Algorithms for Online Content Filtering or Moderation."

The wide range of speech that falls within this broad category includes hate speech, profanity, personal attacks, sleights, defamatory claims, bullying and harassment.[35]

These social media platforms illustrate the usefulness of AI in processing and filtering content. The ensuing question is whether the usage of this AI would reduce, to a certain extent, the liability that the platform may face for content posted by others. Would the use of a smart system mean the platform provider exercises editorial content and hence is liable as a publisher and not as a mere distributor? The article now moves to consider the Malaysian court decision on the liability of online news media for commentaries posted by viewers moderated by a smart system in the Federal Court decision of *Peguam Negara Malaysia v Mkini Dotcom Sdn Bhd & Anor* [2021] 2 MLJ 703.

A review of Mkini case

The brief facts of the case are as follows. Mkini Dotcom Sdn Bhd ('the first respondent') and its editor-in-chief ('the second respondent') operated an online news portal known as 'Malaysiakini'. The portal receives a huge volume of online readership globally, and at times was seen to publish news of sensational nature. On reading a news page, a reader is not required to sign-up or have a subscription to Malaysiakini. However, netizens who wish to leave their comments on any news item are required to have an active paid subscription to the portal. This arguably allows the portal to determine the real identity of netizens and avoid anonymous comments. Mere reading of the published comments, however, does not require any user sign-in. Hence, any readers may view the comments made by registered subscribers as they were available for public access.

Due to the enforcement of the movement control order to control the spread of COVID-19, the courts have been closed for some time. Upon the national shift towards the recovery movement control order, the public service shall resume its operation and the same is with the

[35] Robert Gorwa, Reuben Binns, and Christian Katzenbach, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance," *Big Data & Society* 7, no. 1 (January 1, 2020): 2053951719897945, https://doi.org/10.1177/2053951719897945., accessed December 1, 2022.

judiciary as has been announced by the Chief Justice of Malaysia. On June 9th, 2020, Malaysiakini published an article entitled 'CJ orders all courts to be fully operational from July 1'. The news was arguably reported according to the facts, but it was the comments made by subscribers that caused problems for Mkini as they were found to be scandalous and contemptuous of the Malaysian judiciary. Accordingly, leave was granted to the Attorney General of Malaysia to initiate a committal proceeding against Mkini for the publication of the impugned comments by its subscribers.

There were several issues that were central to this case, including (a) have the respondents rebutted the presumption of publication under s 114A of the Evidence Act? (b) does the 'publication' require the element of intention and/or knowledge to be fulfilled? And (c) did the first and/or second respondents possess the requisite 'intention to publish' for the purposes of scandalising the court contempt?[36]

The first issue concerns the publication of the impugned comments on MKini's portal where reference to Section 114A of the Evidence Act 1950 was made. The applicant submitted that the respondent MKini has facilitated the publication of the impugned comments on its portal and thereby a *prima facie* presumption of publication should arise under Section 114A of the Evidence Act 1950. Accordingly, Section 114A imposes a rebuttable presumption of publication onto anyone whose name appears on the said publication "depicting himself as the owner, host, administrator, editor or sub-editor, or who in any manner". The applicant further submitted that an intention to publish the impugned comments on the part of the respondents was unnecessary to be established.

The respondents raised three main points in their defence. Firstly, the respondents submitted that they should not be responsible for facilitating the publication of the impugned comments due to lack of knowledge as the comments were not created by them. On the second point, Mkini argued that the provisions under the Content Code do not oblige Code subjects to monitor online activities of netizens, unless being prompted by complaints etc. The respondents in their third point submitted to have taken additional measures as follows: 1) by having in-house terms and conditions to warn subscribers against

[36] Peguam Negara Malaysia v Mkini Dotcom Sdn Bhd & Anor [2021] 2 MLJ at p.703.

making illegal and harmful comments; 2) installation of web filter to automatically filter out any bad language found in comments and 3) establishment of an online peer reporting system whereby upon receipt of complaints from any user, the system shall trigger the editor for content moderation process.

To determine whether the presumption of publication under Section 114A raised by the applicant has been rebutted, the Federal Court considered the rebuttals raised by the respondents. Among others, respondents were focusing on the lack of knowledge as a reason to deny liability for the publication of the impugned comments. The Federal Court, after assessing the Hansard of the Parliament and on the legal principles that knowledge can be inferred from facts, held that the respondents were an established company with an appropriate editorial team.[37] The Court also questioned how the editorial team can miss the impugned comments and rejected the view that they do not have knowledge about the comments. On this note, the Court found that the respondents cannot stand on mere denials and "sheer volume could not be the basis for claiming lack of knowledge, to shirk from its responsibility".[38]

Further, the Federal Court also responded to the three measures taken by Mkini as highlighted above. Accordingly, the Court found that the three measures were inefficient to ensure that the impugned comments be removed. On the point raised that the respondent was under no duty to perform active monitoring of content under the Content Code, the Federal Court viewed that the bigger picture had been missed. That the underlying objective of the Content Code was to ensure that all Internet users avoid publishing illegal and harmful content. Although MKini tried to use the Content Code as a defence, this was held to be futile. The Court noted that "the first respondent must have in place a system that was capable of detecting and rapidly remove offensive comments".[39] Accordingly, the respondents failed to rebut the presumption raised under Section 114A and the Court sentenced Mkini to a fine of RM500,000. The Federal Court held that the application for a committal order against the first respondent was allowed and a fine of RM500,000 was imposed on the first respondent.

[37] [2021] 2 MLJ at p.682.

[38] Peguam Negara Malaysia v Mkini Dotcom Sdn Bhd & Anor [2021] 2 MLJ at p.674.

[39] [2021] 2 MLJ at p.674.

Section 114A(1) however could not be extended to the second respondent because there was no evidence that he was at all relevant times named as the owner, host, or editor of the online news portal owned by the first respondent; and there was no evidence that he had the sole discretion to edit or remove any third-party comments. In light of this, the Court was not persuaded that the second respondent was guilty beyond reasonable doubt. As asserted by the applicant, the second defendant was not guilty of contempt.

Does Section 98 of CMA provide a false hope to Code subjects?

Although Section 98 of the CMA is not a new law, no case has yet been brought to court to challenge it. The subsequent issue is whether a content creator must be a registered Code subject? If the answer is in the negative, then compliance with the Content Code is mandatory to the extent that it can provide any sort of protection to the content creators.

Sections 96 and 97 of the CMA establish the Content Code as a self-regulatory guideline for the communications and multimedia industry in Malaysia. Among others, the Preamble to the Content Code highlights that its objectives are to ensure that the "agreed standards of behaviour in respect of industry members" are established. Among the agreed standards of behaviour mentioned in the Preamble include: i) "Promote a civil society where information-based services will provide the basis of continuing enhancements to quality of work and life"; and ii) "The principle of ensuring that Content shall not be indecent, obscene, false, menacing or offensive shall be observed".

Despite its founding roots being in the statutory provisions of the CMA, the wording of Section 98 may be of interest. In its exact words:

> "(1) Subject to section 99, compliance with a registered voluntary industry code shall not be mandatory.
>
> (2) Compliance with a registered voluntary industry code shall be a defence against any prosecution, action or proceeding of any nature, whether in a court or otherwise, taken against a person (who is subject to the voluntary industry code) regarding a matter dealt with in that code."

The provision is not one that is difficult to interpret. It is plain as it intends to say – that registration to become a Code subject is not mandatory for everyone. Under Part 1, Section 4.2 of the Content

Code, states that the Code shall apply to all Content Application Service Providers and but is not limited to: (a) Each member of the forum[40]; (b) Each person who has submitted their agreement to the Forum that they will be bound by this Code; and (c) Each person whom the Commission has directed it in accordance with Section 99 of the Act. A glance through the above groups will indicate that the majority of those who operate in the broadcasting and networking sectors will be required to register as Code subjects. However, an ordinary content creator or Youtuber has the option of whether to become a member or not. This is in line with the spirit of self-regulation as promulgated in Section 123 of the CMA – that participation in the regulatory scheme is usually voluntary.[41]

Given the fact that registration is only mandatory for selected groups of service providers, usually technical in nature, the next question to consider is whether complying with the principles of the Code will accord them any sort of legal protection?

Since registration to be a Code subject is not mandatory for all providers, a non-registrant should, technically, be capable of invoking the protection for hosting third-party content under the CMA. It is clear from the language of Section 98 that compliance with the Content Code shall be a defence against any legal proceeding that relates to the matters dealt with by the Code. From the case review of Mkini, it may be observed that Mkini attempted to use Section 98 of the CMA as a defence. The Federal Court however opined that such was an erroneous way of attempting to seek refuge in the name of Section 98 since Mkini failed to appreciate the general objectives of the Content Code holistically i.e to prevent the spread of illegal and harmful online content.

It is interesting how the Federal Court in that case evaluates who should be entitled to the protection under Section 98. Parties intending

[40] For a list of registered Content Code members, see Content Forum, "CMCF Members," Content Forum, 2021, accessed December 1, 2022, https://contentforum.my/cmcf-members/.

[41] Ian Bartle and Peter Vass, "Self-Regulation and the Regulatory State: A Survey of Policy and Practice" (United Kingdom, 2005), accessed December 1, 2022, http://www.bath.ac.uk/management/cri/pubpdf/Research_Reports/17_Bartle_Vass.pdf.

to seek protection under that section must ensure that they have taken all necessary and expedient measures (such as removal of illegal content), and any of such efforts must be in line with the objectives of the Content Code. The implicit message also goes in line with the equitable principles of 'who comes into equity must come with clean hands'. In the case of *Mkini*, the three measures taken were held to be insufficient hence presumably undeserving to be awarded protection under Section 98.

From the approach taken by the Federal Court in assessing whether Section 98 can become a defence to Mkini, one may ponder to what extent a content creator must go to ensure that the countermeasures taken are in line with the objectives of the Content Code to qualify for Section 98 protection? The Federal Court also differentiated Twitter and Facebook from that of Mkini – the latter being clearly mere conduits[42]. Unlike Twitter and Facebook, Mkini was found to have control over who can post comments on their platform hence was expected to take additional self-control and countermeasures. In summary, Section 98 cannot be read in silo and those intending to seek its protection must prove to the court that all necessary countermeasures to remove prohibited content have been exhausted. At the same time, content creators must prove that they have absolutely no control over third-party content being fed onto their websites. Any efforts taken, be it content takedown, moderation or even using artificial intelligence, must be in harmony with the objectives of the Content Code.

An evaluation of *Bunt* vs *Delfi* vs *Godfrey*: Are we moving backwards?

The case of *Bunt v Tilley*[43] has also been carefully analysed by the Federal Court. In brief, the claimant, Bunt sued three defendants for libel and harassment for allegedly defamatory statements made on Internet chatrooms. Their respective Internet Service Providers were also named as the fourth, fifth and sixth defendants – namely AOL, BT and Tiscali. It was alleged by the plaintiff that as the ISPs gave their respective consumers connections to the Internet, they should be

[42] [2021] 2 MLJ at p.687-688.
[43] [2007] 1 WLR 1243

accountable for the posts complained of. The ISP Defendants moved to have the claim struck out and/or for summary judgement on it.

On determining the extent of liability, the words of Eady J. have been of assistance to Federal Court's analysis as follows:

> "In determining responsibility for publication in the context of the law of defamation, it seems to me to be important to focus on what the person did, or failed to do, in the chain of communication. It is clear that the state of a defendant's knowledge can be an important factor. If a person knowingly permits another to communicate information which is defamatory, when there would be an opportunity to prevent the publication, there would seem to be no reason in principle why liability should not accrue. So too, if the true position were that the applicants had been (in the Claimant's words) responsible for "corporate sponsorship and approval of their illegal activities".[44]

Further in para 23 of Justice Eady's judgment,

> "it is not always necessary to be aware of the defamatory content, still less of its legal significance ... for a person to be held responsible there must be knowing involvement in the process of publication of the relevant words. It is not enough that a person merely plays a passive instrumental role in the process".

The High Court held that all claims against the ISPs were struck out as they play no role in the publication of the impugned communications. The court held that an ISP that performed no more than a passive role in facilitating postings on the internet and did not host the relevant website was not deemed to be a publisher at common law any more than a telephone company would be liable for defamation over the telephone.

It is on this basis that the case of *Mkini* differs from *Bunt*, whereby the liability accrued to Mkini since it knowingly permitted the subscribers to communicate contemptuous communication and that they had the opportunity to prevent such publication but failed to take adequate actions. What Mkini lacked is the element of knowing involvement to be liable for the content. The 2015 case of *Delfi AS v. Estonia* (2015) ECtHR 64669/09 may be equated to the situation of Mkini. Delfi was also an online news portal, with an extensive online readership, similar to Mkini. It published an article that was ruled by

[44] Bunt v Tilley & Others [2006] EWHC 407 (QB), at para 21.

Estonian courts to be defamatory. Upon appeal to the European Court of Human Rights, the court agreed and upheld the decisions of the Estonian court. What was central and relevant in this context is, although there was an argument that Delfi be given immunity purportedly in line with the global legal position on intermediary liability, the European Convention on Human Rights (ECHR) opined that Estonian courts' decision to hold Delfi liable for defamation was within the requirements of necessary and expedient under the ECHR. In this case, the ECtHR also agreed that Delfi was to be treated as a publisher hence reaffirming the findings of the Estonian courts that it was liable for the defamatory comments. In this situation, what transpired was that Delfi was not treated by the courts as an intermediary but as a publisher of the defamatory content – a different position altogether from that of an ISP.

On the other hand, the decision in *Godfrey v Demon Internet Ltd* has taken a different route from *Bunt*.[45] Godfrey sued Demon Internet, an ISP, for defamatory newsgroup posting made available from D's newsgroup servers. Laurence Godfrey, a British lecturer said that an anonymous Internet user published an indecent and defamatory posting and fraudulently ascribed its authorship to him. The comment was made on an online public forum managed by Demon Internet Limited, a UK-based ISP, which did not remove the posting for more than 20 days until its expiration date on the public forum. Subsequently, Godfrey initiated a libel case against the ISP, demanding damages for the claimed defamatory comment. The High Court ruled that the ISP knew or had cause to know that the impugned statement was defamatory as the plaintiff had alerted the firm that he was not the genuine author of the remark. Yet, the Defendants opted not to delete the defamatory message. Accordingly, the Court held that the ISP cannot rely on a viable defence under Section 1 of the UK Defamation Act. It was held that an ISP was a publisher at common law of the defamatory comments posted on the site by an unknown user or the situation is analogous to that of secretary of a golf club which allowed a defamatory statement to remain on a notice board in *Byrne v Deane*.[46]

The position in Godfrey has caused ISPs in the UK to begin removing defamatory materials upon receipt of complaints. This

[45] [1999] EWHC QB 240.
[46] [1937] 1 KB 818 2 ALL ER 204

arguably could lead to chilling of free speech and unwarranted content removal as well as privatised censorship. If *Bunt* and *Godrey* were analysed, one can see that the role played by intermediaries will determine to what extent their liability should be – whether active or passive role. The position taken in *Godfrey* has been avoided in *Bunt* and other recent decisions, whereby ISP or telephone company who plays a passive role in communicating electronic messages should to a certain extent, enjoy a degree of immunity. And similarly, the decision in *Bunt* and *Delfi* has also been accepted in *Mkini* case, but differed because how Mkini behaved was not comparable to an ISP in Bunt. Hence in *Mkini*, it is submitted that the Federal Court was expecting an intermediary to play a far more passive role to avoid liability for publication of illegal content.

How should intermediaries act after this decision?

Given the development of recent cases analysed above, one can draw an early conclusion that the legal framework of intermediary liability is far from certain. With regards to copyright, an ISP will enjoy a safe harbour once it takes action upon notice of the infringing site. The position with regard to defamatory, seditious or contemptuous content is stricter, even though the content may be posted by a third party. The intermediaries must do more to avoid liability as all parties in the chain of communication are strictly liable upon knowledge. The usage of any kind of moderation and filtering may sway the finding to liability as the intermediary consciously act as an editor. The liability of an editor is coterminous to publisher. Thus, if the intermediary wants to remain as an intermediary, it should not do anything towards the content, thus, able to claim as a passive 'middle-man.'

The examples cited above suggest, *inter alia*, that payment network systems, ISPs and hosting providers should be called intermediaries, and enjoy immunity as far as liability for the publication of third-party content on their platforms. In a situation where a payment network system provider provides an online banking service to a client, for example, to wire money from one account to another, then one can safely say that it merely acts as a passive intermediary for such service. Nevertheless, when a payment network provider creates content on its website, then the position shifts to a content creator who shall be answerable to any content that appears on such a website.

Taking into consideration the situation in *Mkini* and *Delfi*, the line between an intermediary that enjoys immunity from that who indulges in the risks of a third-party publication is far from clear. It all depends on the level of engagement as far as content creation, and editorial control are concerned. In the above example of a payment network provider, if at the same time it owns a social media page that enables comment features, the editor needs to be cautious of what liability it is capable to attract for such third-party content, ranging from defamatory, hatred and illegal remarks.

Depending on the law in each jurisdiction, once an intermediary indulges in a semi-passive or active role in creating content of a user-generated nature, they must be presumed to know of the worst thing that can happen and take all necessary initiatives to avoid them. Learning from *Mkini* and *Delfi*, a passive effort of waiting to be notified upon complaints and usage of filters were held to be inadequate, if one were to claim that they have done enough to filter out illegal content.

RECOMMENDATION AND CONCLUSION

The introduction of notice and takedown procedure to minimise the liability of intermediaries for third party content does not eliminate all potential liabilities. Case laws around the world demonstrate that intermediary liabilities differ according to types of content i.e., copyright infringing, defamatory, sedition, sacrilegious, pornography etc. The inconsistent stand depends substantially on the severity of the offences and how the law traditionally treats third parties' accountability in the conduct. Copyright infringement has traditionally been deemed to be a strict liability and a third party's involvement would equally be culpable. On the same token, for defamatory or sacrilegious content, the court does not distinguish between publisher, editor and distributor's liability if there is knowledge of the culpability of the conduct.

An intermediary can potentially be liable for a third party's content if there is proof that the intermediary has the ability to control, choose or edit content. Mkini has proven that with regard to contemptuous content, the same strict liability stand has been taken by the court. One major flaw with the approach taken by the court in *Mkini* is the inability to distinguish between filtering 'harsh words' with content contemptuous of the court. The court did not venture further to

determine whether the filtering software adopted by Mkini was actually capable of filtering content critical of the court. In doing so, the Court has practically cast a ruling that all publishers or online platforms must stay away from commenting on current issues that pertain to the conduct of the administration of justice as that would be deemed to be 'contempt of court'. With due respect, we argue that such a stand may ignore the fundamental right to express an opinion on topical issues of public interest.

In the final analysis, one must be mindful of the report done by the Transatlantic Working Group where it was conceded that: "AI" however, is not a simple solution or a single type of technology" – that there are various forms of AI and automation used in content moderation and that the existing content curation focussing on hate speech, violent extremism and disinformation varies greatly depending on the technology used. Among the report's key recommendations is that- automation in content moderation should not be mandated in law because the state of the art is neither reliable nor effective.[47] Most importantly, the Report concedes that the context of the message is more important than the words used and often this is ignored in algorithm-based content moderation.[48] The Report specifically alluded to factors not taken into the AI system such as history, politics, and cultural context. The Report further espouses the view that 'intermediary liability laws should neither mandate, nor condition liability protection on, the use of filters.[49]

Despite the advances in content moderation and the truism that the system is not infallible, it remains to be seen whether content contemptuous of courts based on its subjectivity could easily be contained through algorithms. Unlike profanity, obscenity, hate speech, and sexually explicit materials that could easily be 'targeted' by content moderation, content that criticises a court judgement whilst the litigation is still ongoing is local-centric and could not easily be identified, monitored or tracked and managed by the system. In this

[47] Emma Llansó et al., "Artificial Intelligence, Content Moderation, and Freedom of Expression," Transatlantic Working Group, 2020.

[48] Llansó et al.

[49] Emma Llansó et al., "Artificial Intelligence, Content Moderation, and Freedom of Expression," Transatlantic Working Group, 2020, accessed December 1, 2022, https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

context, we have to be mindful of the fact, despite the aggressive use of AI content moderation by big platform providers, the technology is nowhere near perfect. Intermediaries that choose to deploy some form of content moderation should not simply be seen as performing the publisher nor the editor role. The widespread disinformation, misinformation, toxic speech, profanity, and obscenity warrant the big players to exercise due care by playing a role in stemming such illegal and harmful content from being disseminated online.

ACKNOWLEDGMENT

The authors extend our deepest appreciation to the Ministry of Higher Education Malaysia for funding this paper and related research work via Fundamental Research Grant Scheme FRGS/1/2021/SSI0/UIAM/01/1.