THREE

How Online Defamation Cases Are Decided

This chapter will rely on the *data-dissemination* scenarios outlined in Chapter 2 in order to guide its analysis into the difficulties potential claimants face in cases of online defamation. It is separated into two parts; Part I is primarily concerned with the impact of the Defamation Act 2013 and how this has negatively affected claimants in online defamation actions. In this sense, Part I focuses on *legal* changes and the challenges these pose to those defamed in various ways on the web. Part II of this chapter is concerned instead with *technological* changes, which have increased the prevalence of defamatory content online and altered the digital landscape, and the intersection of these changes with libel actions. The overarching theme of this chapter is an examination of how English defamation law and the judiciary have responded to the changing digital landscape and increasing defamatory content online. This chapter will argue that there is much more that needs to be done in order to reinstate reputational interests in an online context and that the law as it currently stands is failing claimants in a number of ways.

★★★

Part I: Difficulties for claimants posed by the Defamation Act 2013

This section will firstly consider the s 1 threshold of 'serious harm' as caused or likely to be caused by a particular statement required by the 2013 reform. This reform was introduced by the Act with the intention of making bringing an action in defamation by a claimant more difficult and therefore 'weeding out' what were considered by some to be weaker claims. As will be argued, s 1 has a potentially prohibitive impact on everyone who wishes to mount an action in defamation by raising the threshold to bring an action across the board – but may pose particular problems to those arguing that they have been defamed on the internet. This will invoke the *defamation by social media* scenario outlined in Chapter 2. Secondly, the introduction of a limitation period of one year for repetitions of defamatory statements introduced by s 8 of the Act will be considered. This replaced the long-standing rule in *Brunswick v Harmer*,[1] where every repetition of a defamatory statement gave rise to a fresh action in defamation, regardless of how much time had passed between this and the initial publication. It seems clear that this limitation period will also apply to defamatory statements posted on the internet – and invokes the *repetition of statements online over a year later* scenario, also outlined in Chapter 2.

Both of these changes introduced by the 2013 reform have 'swung' English defamation law in favour of freedom of expression, at the expense of the protection of individual reputation and therefore personal dignity. These legal changes and both defamation scenarios will be examined in the context of defamation through the mode of the internet.

[1] *Duke of Brunswick v Harmer* [1950] 175 ER 441: the traditional rule was that every republication of a defamatory statement gives rise to a new claim.

I. Online publication and the 'serious harm' threshold

a. Background to the reform

Libel reform in the Defamation Act 2013 could not have come at a worse time. As discussed in Chapter 1, the world is now in the grip of the digital age – with increased threats to individual reputation through defamation on the internet. Despite this, English and Welsh libel law was changed in 2013 in order to tilt legal precedent in favour of *expression* rather than the protection of individual reputation. One of the key ways this rebalancing was achieved through the 2013 Act was by raising the threshold a claimant must meet in order to bring an action in defamation, irrespective of whether the defamatory statement appears online or in print. This provides a further barrier to the protection of personal dignity through reputation rights in the face of the unbridled potential to defame using the internet. The purpose of the introduction of the 'serious harm threshold' in s 1 of the Defamation Act 2013 was to 'raise the bar' as to what type of statement could be actionable in defamation law, to include only those which have caused (or could likely cause) serious reputational harm to the claimant.[2] The idea behind this was that it would curtail potentially 'spurious' claims and strengthen freedom of expression generally. The High Court in *Courtney v Ronksley* explains the change in the law:

> Defamation is an abridgment of free speech. When it introduced the serious harm test, Parliament's intention was to allow a greater margin to free speech, and to prevent the scarce and precious public resources of the senior courts from being occupied with defamation challenges to others' freedom of expression, unless

[2] Section 1(1) Defamation Act 2013.

> objectively demonstrable real-life reputational impact can be established, on ordinary causational grounds, and to a proper threshold of gravity.[3]

The 2013 reform was enacted because journalists, human rights campaigners, scientists and other pressure groups argued that their legitimate speech was being curtailed as a result of concerns about libel actions. As a result, it became an election promise on both sides of the House that English libel law would be reformed to reprioritize expression interests.[4] This was partly achieved by the introduction of the serious harm threshold in s 1. Sewell has argued that the introduction of s 1 (and its interpretation by courts) is a good thing, as it is 'inspiring parties to resolve their case away from court'.[5] The accuracy of this can be questioned; certainly, if one considers the number of high-profile celebrity defamation cases in the last five years this does not seem to be true – but the sentiment that this in fact would be a positive outcome regardless should also be challenged.[6] If there has been reputational damage arguably accrued on the basis of a defamatory statement (that

---

3 *Courtney v Ronksley* [2024] EWHC 572 (KB) [64].

4 Charlie Sewell, 'More serious harm than good? An empirical observation and analysis of the effects of the serious harm requirement in section 1(1) of the Defamation Act 2013' (2020) 12(1) *Journal of Media Law* 47, 50–1 and English PEN, 'Libel Reform Campaign' www.englishpen.org/campaign/uk-free-speech/libel-reform-campaign/ accessed 28 November 2024. Also see Andrew Scott, 'Impact case study: Reforming England's libel law' www.lse.ac.uk/Research/research-impact-case-studies/reforming-englands-libel-law#:~:text=He%20found%20that%20reform%20of,philosophy%20and%20human%20rights%20law accessed 28 November 2024.

5 Sewell (n 4) 56.

6 Such as *Vardy v Rooney* [2021] EWHC 1888 (QB); *Depp II v News Group Newspapers Ltd* [2020] EWHC 2911 (QB); and *Blake v Fox* [2023] EWCA Civ 1000.

cannot be shown to be true),[7] the correct method to rectify the damage to personal dignity caused is for claimants to 'have their day in court' and to have it proclaimed to the world in a legal decision that a defendant is liable for the statement. If claimants are dissuaded from pursuing potentially legitimate actions on the basis of s 1, this may in fact operate as a barrier to justice. Sewell argues that it will at the very least deter 'opportunistic claims',[8] although there is little evidence that this was a problem to begin with. To the contrary, with the ease, speed and worldwide reach of the internet it is now simple to disseminate defamatory information about another, particularly using social media websites (such as the case in the *defamation by social media* scenario).

Section 1 was introduced despite the fact that safeguards for freedom of expression had been present at common law for some time before the adoption of the Defamation Act 2013. The pre-existing case law of *Thornton* and *Jameel* dictated that a 'tendency'[9] to cause 'substantial' reputational harm was the threshold necessary to bring a claim,[10] a threshold articulated to account specifically for Article 10 European Convention on Human Rights (ECHR) concerns. Hyde has noted that spurious claims were therefore already able to be struck out

[7] According to s 2 Defamation Act 2013, now known as the truth defence (formerly justification) – substantial truth is the threshold required under s 2(1).

[8] Sewell (n 4) 57.

[9] See Iain Wilson and Tom Double, 'Business as usual? The Court of Appeal considers the threshold for bringing a libel claim in *Lachaux v Independent Print Ltd*' (*Inforrm's Blog*, 16 September 2017) https://inforrm.org/2017/09/16/business-as-usual-the-court-of-appeal-considers-the-threshold-for-bringing-a-libel-claim-in-lachaux-v-independent-print-ltd-iain-wilson-and-tom-double/ accessed 28 November 2024.

[10] *Thornton v Telegraph Media Group* [2010] EWHC 1414 (QB) [94] and *Jameel (Youssef) v Dow Jones & Co Inc* [2005] QB 946 [40] and [55].

under common law rules before s 1 was adopted, serving to raise this threshold higher still.[11]

b. Interpretive difficulties: what exactly is the new s 1 'serious harm' threshold?

For a claimant to bring an action in defamation in respect of a post on the internet (such as on Facebook) or otherwise, s 1(1) of the Defamation Act 2013 tells us that 'a statement is not defamatory unless its publication has caused or is likely to cause serious harm to the reputation of the claimant'.[12] In essence, this challenges a claimant's ability to prove they have an 'evidential basis' for their claim from the outset.[13] It also erodes the distinction between libel and slander – slander is traditionally only actionable with proof of special damage,[14] whereas libel was actionable per se because of its more permanent nature.[15] This distinction, as it applies to online posts (or statements more generally),[16] has therefore ceased to be important – the question now is, how precisely must one adduce whether a post on a website such as Facebook meets the s 1 threshold. Any requiem for the distinction between libel and slander should perhaps be none too hasty, as Van Veeder has observed the distinction was a result of a historical anachronism and difficult to reconcile with modern media such as the internet.[17]

---

11 Richard Hyde, 'Procedural control and the proper balance between public and private interests in defamation claims' (2014) 6(1) *Journal of Media Law* 47, 66.

12 Section 1(1) Defamation Act 2013.

13 Hyde (n 11) 66.

14 *Roberts v Roberts* [1864] 33 LJ QB 249.

15 See, for example, *Monson v Tussauds Ltd* [1894] 1 QB 671.

16 See *Tamiz v Google* [2013] EWCA Civ 68.

17 Van Vechten Veeder, 'The history and theory of the law of defamation II' (1904) 4(1) *Columbia Law Review* 33, 54

Perhaps because of the rather thin explanation of the new threshold in s 1(1) of the Act, the question of what precisely the new test requires has been left to the courts,[18] and the idea of showing 'likelihood' of serious harm occurring has resulted in different interpretations.[19] The very early case of *Cooke* held that s 1 operated a higher standard than what was previously required at common law, although evidence will not always be required.[20] Similarly, another early case, *Ames*, held that s 1 had raised the threshold as previously existent at common law, although serious harm could be inferred without evidence of an adverse reaction from readers.[21] A few years later in 2017, in *Monroe v Hopkins*, part of the claimant's case with regards to what constituted serious harm in s 1(1) was that she had 'trouble sleeping' and was anxious about the backlash she could face on social media due to defamatory tweets about her, an argument accepted by Mr Justice Warby.[22] The learned judge went on to say that 'the serious harm requirement is satisfied, on the straightforward basis that the tweets complained of have a tendency to cause harm to this claimant's reputation in the eyes of third parties, of a kind that would be serious for her',[23] implying that the threshold is not only highly practical in nature, but also a somewhat subjective standard. Mr Justice Warby went on to observe that that if the statement is seriously defamatory and widely published, this may be enough to satisfy s 1 without any further evidence being adduced.[24]

After a number of years of uncertainty for claimants, the judgment of the Court of Appeal in *Lachaux* was delivered in 2017.[25] The Court of Appeal famously disagreed with the

[18] Sewell (n 4) 52.
[19] Ibid.
[20] *Cooke v MGN Ltd* [2014] EWHC 2831 (QB) [37] and [43] respectively.
[21] *Ames v Spamhaus Project Ltd* [2015] EWHC 127 (QB) [55].
[22] *Monroe v Hopkins* [2017] EWHC 433 (QB) [64].
[23] Ibid [70].
[24] Ibid [69].
[25] *Lachaux v Independent Print Ltd and Another* [2017] EWCA Civ 1334.

decision of (once again) Mr Justice Warby in the same matter at the High Court. The latter had argued previously that the bar had been raised from the common law position and that claimants must go further than merely showing the *tendency* to cause reputational harm. Judge Warby argued that it must be shown by evidence that serious harm to one's reputation had been caused, or was likely to be caused in the future.[26] The Court of Appeal did not adopt Judge Warby's reasoning – rather, they found that the threshold had merely been changed from that which was *substantial* to *serious*, with seriousness nevertheless being a more significant standard in terms of proof required to meet it.[27] The Court of Appeal found that Judge Warby's more radical interpretation, which was a more marked change from the position at common law, was not what parliament had intended.[28] Essentially, the position the Court of Appeal adopted in *Lachaux* was that of '*Thornton*-plus';[29] the threshold was raised from the position at common law, but raised only slightly. The decision was once again appealed, this time in the Supreme Court, where the debate about how high the bar had exactly been raised from the position at common law was firmly laid to rest. Lord Sumption delivered the judgment, rejecting the Court of Appeal's interpretation of s 1 in favour of Judge Warby's original decision at the High Court.[30] Lord Sumption held that 'actual facts' about the impact of the words are relevant to determining s 1 and it 'raises the threshold of seriousness' above that which existed in *Jameel* and *Thornton*.[31] Whether s 1 was met, Lord Sumption found, 'depends on a combination of the *inherent tendency* of the words and *their actual impact* on those to whom they were

[26] See Wilson and Double (n 9).
[27] Ibid. *Lachaux v Independent Print* (n 25) [56ff].
[28] *Lachaux v Independent Print* (n 25) [56]–[59].
[29] Sewell (n 4) 55.
[30] *Lachaux v Independent Print* [2019] UKSC 27 [20].
[31] Ibid [12].

communicated'.[32] The Supreme Court decision signalled a considerable change from the position at common law – the court found that s 1 primarily concerned 'factual investigation' about the statement's impact.[33] Lord Sumption held that to satisfy s 1 claimants must build an evidential case and adduce evidence that demonstrates that the statement(s) complained of have had a reputationally damaging impact.[34] If this is not possible, claimants may plead an inferential case, which argues that the statements are *likely* to cause serious reputational harm, with reference to a number of factors, such as: the scale of publications, that people who knew the claimant had read the statement(s), that they could come to the attention of others in future and the gravity of the statements themselves.[35] This was a controversial interpretation, as it overturned the long-standing presumption-of-damage principle for libel at common law.[36] This decision is also significant for the purposes of this book as it made clear that the bar had been firmly raised from the common law position; s 1 is now a significant hurdle for claimants to overcome and is most likely to be satisfied with hard evidence about a statement's defamatory impact, such as testimony from individuals who have read a statement and, as a result, thought less of a claimant.[37] This is not always easy to demonstrate, particularly in relation to posts on social media sites like X in the *defamation by social media* scenario, which can be more transient in nature. People reading posts behind their X 'handles' may also be reticent to come forward. This will be discussed in more detail later. Lord Sumption also found that taking into account damage to the claimant's reputation from people reading the statement who *didn't know her at the*

32 Ibid [14]; emphasis added.
33 Ibid [12] and Sewell (n 4) 54.
34 *Lachaux v Independent Print* (n 30) [21].
35 Ibid.
36 Sewell (n 4) 55.
37 *Lachaux v Independent Print* (n 30) [21].

time is also perfectly legitimate in establishing whether s 1 is met.[38] However, once again this is a difficult line of enquiry for a claimant to substantiate, as individuals who did not know them at the time of reading will scarcely know them better *after* reading the statement complained of, and once again may be both difficult to get in touch with on the part of the claimant and even more difficult to convince to come forward in the case of online posts. For the purposes of what a claimant must now prove for an actionable case in defamation, what matters most is how the *Lachaux* 'factors' have been distilled by the case law that has followed. The court in *Coker* summarizes factors that may go towards an *inferential* case: 'serious harm can in principle be proved on a combination of (a) the meaning of the words; (b) the situation of the claimant; (c) the circumstances of publication; and (d) the inherent probabilities.'[39]

However unfortunate for claimants, Lord Sumption was undoubtedly right that parliamentary intention was to 'raise the bar' in s 1, as otherwise enacting it would have been redundant. Indeed, the decision has drawn praise at least for its sound legal interpretation of the 2013 statute.[40] Despite the fact that the decision 'is legally convincing',[41] the new and far stricter interpretation of s 1 has concerning timing given the potentially unlimited defamatory potential posed by the internet, with its ease and multitude of platforms on which defamatory remarks can be spread (including social media, in the *defamation by social media* scenario – but also beyond it). Artificial intelligence (AI) programs and new mediums such as digital worlds, many of which are free to use, present new threats to reputation rights online. As a result, there is a greater

38 Ibid [25].

39 *Coker v Nwakanma* [2021] EWHC 1011 (QB) [12].

40 David Erdos, 'Case comment: Serious harm to reputation rights? Defamation in the Supreme Court' (2019) 78(3) *Cambridge Law Journal* 510, 512.

41 Ibid.

menace to reputation and therefore individual dignity today than ever before and this seems incongruent with parliament's recent intention to make it *more difficult* to bring an early stage claim in defamation by installing this 'control mechanism'.[42] As Erdos puts it: 'it remains the case that untrue and unfair attacks on reputation are increasing (principally online) and defamation law may often not provide an effective avenue for vindicating the rights that are thereby impaired.'[43]

This new threshold has also made case outcomes difficult to predict for claimants. Legal practitioners and academics alike are awaiting more detailed guidance (that will emerge over time) of clear themes that will demonstrate s 1 is likely met.[44] This current uncertainty creates a climate of fear for those considering litigating over a potentially reputationally harmful statement online, which is already a costly process, particularly if a claim is disposed of at an early stage due to this threshold.

c. How does one evidence serious harm caused or likely to be caused by an online post?

As explained earlier, the Supreme Court decision in *Lachaux* has set a high bar. The investigation that is now required by the courts to meet s 1 risks that 'defamation actions will become stuck in lengthy and expensive interlocutory proceedings',[45] as Erdos has observed. Overtly, an assessment of whether s 1 is met by a post, for example, on Facebook, should not be inherently different to argument that it is met by virtue of a publication in print. However, in practice, the way evidence can be adduced to meet this threshold and the type of evidence that can be offered will likely be different for a

[42] Sewell (n 4) 69.
[43] Erdos (n 40) 513.
[44] Sewell (n 4) 70.
[45] Erdos (n 40) 512.

post online than in print – with evidential hurdles a yet higher mountain to claim for those defamed online. A detailed recent case on s 1, *Amersi v Leslie*, noted that in order to meet s 1 the court examines a combination of the *impact the words have had* and the *inherent tendency* of the words,[46] with reference to this sentiment expressed by the Supreme Court in *Lachaux*. The High Court in *Amersi* also made a number of other salient observations. Firstly, even publication (online or otherwise) to a 'relatively small number of publishees may yet cause very serious harm to reputation'[47] – so the question is not merely one of the number of 'hits' online, although this is clearly relevant. On a sympathetic note to claimants, the court observed that in 'mass publication' cases – such as mass dissemination using a social media platform such as X – 'a claimant may struggle to identify, or produce evidence from all those to whom an article was published'.[48] Indeed, this would be an impossible task for some defamed online, where posts on X may have reached hundreds of thousands of followers. The court observed that this is perfectly acceptable – however, in cases where the matter was published to a single person, evidence would be sought and needed.[49] This may prove difficult for claimants in situations where third-party readers have been contacted privately through the 'messenger' function on Facebook with defamatory statements, perhaps by someone they do not know, so they do not necessarily want to come forward and support a claimant who has been defamed. Lord Justice Warby has also observed that in relation to online publication to strangers,

[46] *Amersi v Leslie* [2023] EWHC 1368 (KB) [144] and *Lachaux v Independent Print* (n 30) [14].

[47] *Amersi v Leslie* (n 46) [145]; *Sobrinho v Impresa Publishing SA* [2016] EMLR 12 [47]; *Dhir v Sadler* [2018] 4 WLR 1 [55 (i)]; *Monir v Wood* [2018] EWHC 3525 (QB) [196].

[48] *Amersi v Leslie* (n 46) [146].

[49] Ibid.

a mass of publications could engender the s 1 threshold has met – each individual publication does not have to, in itself, create serious harm to one's reputation.[50] This is clearly a logical and practical approach. However, *Amersi* noted that one cannot 'aggregate' reputational harm for the purposes of the serious harm threshold by mere generalities: 'ultimately a claimant must satisfy s.1 by evidence'[51] in respect of *each individual statement* complained of. Therefore, one cannot argue that a number of posts all relating to the claimant *as a whole* may cause serious harm to their reputation – rather, each statement complained of must be individually examined as to whether it meets this high threshold, as the High Court recently observed in *Goldsmith*.[52] This is certainly a daunting task and, in multiple-post cases concerning online speech, will lead to many posts falling at this hurdle at an early stage. This runs contrary to the fact that it would in fact make contextual sense in some *defamation by social media* cases to run posts together and for s 1 to be examined holistically – as often posts made about the same matter over a short time period are read in succession by online observers. In fact, due to the short character limit of X,[53] a defamatory meaning in fact could be observed by reading several different posts together, rather than separately – as people often label posts in the style of 'post 1 of X number'. This is also the case for short video-sharing platforms such as TikTok, where a user plays a short video clip and this is then followed by another related video that then plays immediately after – or a person can return to an uploader's homepage and click to view the next video adjacent in the viewing

[50] *Banks v Cadwalladr* [2023] EWCA Civ 219 [49].

[51] *Amersi v Leslie* (n 46) [162].

[52] *Goldsmith v Bissett-Powell* [2022] EWHC 1591 (QB) [152].

[53] Which for standard users is 280 characters. See https://x.com/premium/status/1623411400545632256 accessed 29 November 2024.

panel, several of which may piece together a longer, potentially defamatory, narrative.[54]

What has emerged from the case law to date is that the serious harm threshold is an exercise in fact finding, with a claimant expected to provide evidence that requisite reputational damage was caused by the complained-of statements. However, if hard evidence cannot be produced to point towards concrete damage to individual reputation (due perhaps in part to the difficulties as outlined), an inferential case can instead be pled, which argues that by *inference* it is clear that serious harm to reputation – and therefore personal dignity – has been done, or is likely to be done in the future. It is submitted that this is more likely the route to be taken in respect to claimants who are defamed online, as it may be harder to adduce hard evidence in the form of – for example – third-party reader testimony that can corroborate the claim. Unfortunately, arguing a purely inferential case is fraught with challenges and it is often more difficult to prove that the serious harm threshold has been met. In inferential cases, the *Lachaux* factors of 'the meaning of the words, the situation of claimant, the circumstances of publication and the inherent probabilities' are relevant.[55] The courts have powerfully stated that there is a difference between an inferential case and 'speculation',[56] although it is not abundantly clear what this difference is and any decision based entirely on inference is likely to involve some level of speculative thought. Particularly with cases, be they online or offline, that involve a small number of publishees, s 1 has made it extremely difficult to plead a successful inferential

[54] A recording can be up to ten minutes in length if one records using the TikTok function. See Camera Tools, TikTok Support, https://support.tiktok.com/en/using-tiktok/creating-videos/camera-tools accessed 29 November 2024.

[55] *Blake v Fox* [2024] EWHC 146 (KB) [50] and *Lachaux v Independent Print* (n 30) [21].

[56] *Sivananthan v Vasikaran* [2023] EMLR 7 [53].

case without hard evidence of seriously harmful defamatory impact – as the 'percolation' effects of mass publication cannot be relied on.[57] Finally, it should be noted that the single most important factor that is repeatedly relied on in defamation proceedings going to the weight of the serious harm threshold is the *seriousness of the statement complained of*.[58] This is the case regardless if defamatory allegations are spread using the internet or by other mediums. On a number of occasions, the courts have gone so far as to abandon substantive analysis of the s 1 threshold if the statement complained of is sufficiently reputationally grave – the idea being that this threshold has been implicitly met.[59] On this topic, Mr Justice Warby has stated, 'it is certainly not necessary in every case to engage in a detailed forensic examination of the precise factual picture, in order to determine whether the serious harm requirement is satisfied'.[60] An obvious question remains: exactly how serious do allegations have to be to negate the fact-finding exercise of s 1's threshold? The answer, unhelpfully, is as yet unclear – offering little comfort to claimants and defendants alike. Perhaps unsurprisingly, early studies have found that imputations of morally reprehensible conduct, such as sexual impropriety or conduct striking to the core of one's personality, are more likely to result in s 1 as met.[61]

d. What is the significance of viewership and engagement metrics to s 1?

In light of the reform in s 1, the courts have struggled with the question of the precise significance of online viewership or engagement metrics of a defamatory post. Several decisions

[57] *Amersi v Leslie* (n 46) [159].
[58] Sewell (n 4) 58–60.
[59] See, for example, *Coker v Nwakanma* (n 39) [33].
[60] *Monroe v Hopkins* (n 22) [69].
[61] Sewell (n 4) 58–60.

have been at pains to stress that satisfying the threshold 'was never just a "numbers game": one well directed arrow [may] hit the bull's eye of reputation'.[62] Despite this, in reality, many cases that concern *defamation by social media* consider engagement metrics as relevant to the threshold being met. In the recent case of *Bukhari v Bukhari*, which related to a number of posts on X that alleged the claimants were criminals, counsel for the claimant argued that the defendant had nearly 2,000 followers and 'tagged' other individuals in the posts, including one that had millions of followers – in an attempt to satisfy s 1.[63] Counsel also argued that the view count of each post was relevant, including the number of 'likes' and re-posts.[64] Further, in *Hayden v Family Education Trust*, which again concerned a comment on X, the numbers of re-posts, account followers, those who were 'tagged' and how long the post remained on X all were found to be crucial to the assessment of whether the serious harm threshold was met.[65] A total of 3,800 followers on X was deemed to be significant for the purposes of s 1, as was the total extent of the publication on X to 45,000 users.[66] It was also deemed relevant that there was a 'social media and mainstream media frenzy' which coincided, driving people to X – demonstrating that claims in respect to online posts can also be bolstered by traditional media coverage. Importantly, *Hayden* noted a number of things in relation to social media posts and the s 1 threshold. Firstly, the duration of the post is important to determining analytics – if a post is only available for a short time and then deleted, even if there is a large follower base, it will be harder to obtain representative analytics but also to prove a large breadth of publication, as not every

62 *Amersi v Leslie* (n 46) [145] quoting *King v Grundon* [2012] EWHC 2719 (QB) [40].
63 [2023] EWHC 427 (KB) [87].
64 Ibid.
65 *Hayden v Family Education Trust* [2023] EWHC 950 (KB) [2] and [8].
66 Ibid [8].

follower will seen the post in the narrow time period it was available.[67] This poses an additional problem for claimants, as it allows defendants to (perhaps strategically) post a potentially reputationally damaging statement on social media for a short time but then quickly delete it, the swift deletion leading to the serious harm threshold being difficult to prove on the part of a claimant on whom the burden firmly lies.[68] This approach does little to deter defamation online in the digital age. Numbers of social media 'followers' was also seen as strongly relevant to the determination of the s 1 threshold in *Miller v Turner*, where 16 posts on X and a webpage were complained of.[69] There is, then, a clear disconnect here – some members of the judiciary have claimed that decisions about the serious harm threshold are not a 'numbers game', yet social media analytics seem to be playing a pivotal role in how the serious harm threshold is decided in online defamation cases. Using analytics in this way is a somewhat reductive exercise and it is hard to believe that this is what parliament exactly envisaged when enacting s 1. The court in *Wright v Granath* found that expert evidence about how many people read a tweet is not expected or necessary,[70] although it is noted here that given a judicial lack of technological expertise in certain cases it might be – notwithstanding the time and economic cost that comes with it.

However, not all decisions regarding the serious harm threshold and posts on social media have rendered problematic results for claimants – certain judges have led the way with practical and technologically insightful decisions. Several cases have correctly observed that *readership* numbers are often larger than engagement numbers of an online post, as no account

67 Ibid [34].
68 Ibid [43].
69 *Miller, Power v Turner* [2023] EWHC 2799 (KB) [47].
70 *Wright v Granath* [2022] EWHC 1181 (QB) [59].

is required to read posts on Twitter/X.[71] The court in *Riley* observed that tweets can also be subject to a 'percolation' effect (otherwise known as the 'grapevine' effect) that inflates estimated readership numbers beyond those which pure analytics derived from social media websites suggest.[72] This takes into account the popularity or craze that surrounds certain online disputes and posts – which generates more readership traffic to these types of statements online than can be reflected in 'pure' analytics. Certain decisions have also taken into account whether a defendant's social media account was growing at the time of publication (and therefore readership wider than usual).[73] In a logical approach, the court in *Packham* found that replies to tweets can help to evidence serious harm as met,[74] as replies represent the tweets' impact on the readership without necessitating individual testimony in court. A more pragmatic approach than pure analytics alone was also suggested in the recent case of *Blake v Fox*.[75] After considering the inferential case to s 1, the court then looked at the actual *state of affairs on the ground at the time*: what people on X were saying and likely thought of Laurence Fox, if people actually believed that his statement was a rhetorical device and what newspapers were writing, in order to establish if serious harm had been caused to the claimant's reputation.[76] Influence of the particular defendant's opinions (on social media) in a particular nexus of society has also been held to be relevant to establishing s 1.[77] These are all logical factors for consideration rooted in the practicalities of proving the serious harm threshold as met for the purposes of online defamation. It is hoped that more

[71] If the profile is not restricted, or 'private'. Ibid [38(b)].
[72] *Riley v Murray* [2022] EMLR 8.
[73] *Wright v Granath* (n 70) [54].
[74] *Packham v Wightman* [2023] EWHC 1256 (KB) [74].
[75] *Blake v Fox* (n 55) [50].
[76] Ibid [79]–[103].
[77] *Miller, Power v Turner* (n 69) [47].

decisions in this spirit will be handed down by the courts in future and an approach wholly centred around analytics be abandoned.

e. Is there a different approach to s 1 where the internet is concerned?

Aside from different ways that claimants in online defamation cases must currently *plead* s 1 (for example, by adducing corroborating social media metrics in the *defamation by social media* scenario), there is also some evidence to suggest that the judiciary *treat* defamation on the internet differently to defamation through traditional mediums. Most famously, the Supreme Court in *Stocker v Stocker* found that Facebook: '[is] a casual medium; it is in the nature of conversation rather than carefully chosen expression; and that it is pre-eminently one in which the reader reads and passes on.'[78]

This author has argued elsewhere that this was an unfortunate choice of words on the part of Lord Kerr, as the (perhaps unintentional) implication was that social media websites are inherently transitory mediums – and therefore should be taken less seriously in terms of defamatory potential by the courts. Today, this is clearly out of step with the realities of modern media. Many people now reverently and earnestly read news through social media websites and take certain comments posted on platforms such as Facebook and X extremely seriously.[79] As the present author has argued elsewhere,[80]

[78] *Stocker v Stocker* [2019] UKSC 17 [43].

[79] In 2024, Ofcom reported that over half of adults use social media to find out news. Dan Milmo, 'Internet replaces TV as UK's most popular news source for first time' *The Guardian* (10 September 2024) www.theguardian.com/media/article/2024/sep/10/internet-tv-uk-most-popular-news-source-first-time#:~:text=More%20than%20half%20of%20UK,used%20by%2018%25%20of%20adults accessed 2 December 2024.

[80] Fiona Brimblecombe, '*Stocker v Stocker*' in Lewis Graham and Jennifer Russell (eds), *The Supreme Court at 15: Reflections on Private Law* (Routledge 2025).

this statement of Lord Kerr has not proved fatal to online defamation litigation. It has been easy to sidestep any inference of the Supreme Court to this effect by stating that, in the relevant circumstances, the given statement *was* in fact sufficiently serious in impacting the claimant's reputation. The Court of Appeal has also recently explicitly acknowledged in *Blake v Fox* that not all tweets are 'conversational' – with many instead presenting purported facts, designed to be taken seriously.[81] Similarly, in *Miller*, Mrs Justice Collins Rice stated when discussing Twitter/X: 'It tends to be thought of as largely consequence-free for Real Life, and so is sometimes described as a playground. But it has given a new meaning to the word 'troll' in the English language. And online behaviour *can and does have real life consequences*.'[82]

In *Banks v Cadwalladr*, the court distilled a number of overarching factors that can add to the weight of s 1 being met, reminiscent of Lord Nicholls' 'laundry list' in *Reynolds*.[83] Many of these have been discussed earlier in this chapter. For the sake of completeness, an abbreviated version of this list goes as such: s 1 is a higher threshold; each statement must meet s 1 individually; there is no presumption of serious harm; the question is of the 'actual impact' of the statement; the reactions of others are valid; evidence in proving harm is persuasive; 'sometimes inference may be enough'; bad reputation of the claimant is relevant; background context is relevant; extent of publication is relevant and malice is *not* relevant on the part of the defendant.[84] In addition to this already lengthy list, a number of other factors can also be distilled from the early case of *Monroe* that relate specifically to posts on social media and s 1. These are:

[81] *Blake v Fox* n 6 [67].
[82] *Miller, Power v Turner* (n 69) [5]; emphasis added.
[83] *Reynolds v Times Newspapers* [2001] 2 AC 127 [205].
[84] *Banks v Cadwalladr* [2022] EWHC 1417 (QB) [51].

1. If the post is transient, and how often it has been viewed.
2. The 'credibility' of a publisher 'in the eyes of publishees'.
3. Any evidence that 'no harm was done' to the claimant's reputation, based on 'contemporaneous social media activity'.
4. If the claimant suffered a 'torrent of abuse' as a result of the post/s.
5. The claimant's own responses on social media.
6. Any other (non-social media) coverage of the post complained of.[85]

The problem with this list is that it is a rather leading one that is skewed against finding that the s 1 threshold is met by social media posts. The point about transience is problematic as it invokes the need to employ user metrics, which are not only unreliable but also sometimes difficult to find, as discussed earlier. Credibility may also be an additional barrier for online posts to meet, as one would then be forced to argue that the person making the post had considerable social pull or sway – which may not always be possible for claimants, with the post nevertheless reputationally damaging. Similarly, it is hard to understand how a court could accurately adduce if 'no harm was done' to a claimant by other contemporaneous social media activity – as for every supportive message a claimant might receive, many other individuals who now think poorly of the claimant may have stayed silent. Factor 4 seems to suggest that a 'torrent of abuse' is another necessary hurdle that a claimant must overcome, which will not always be provable at that particular moment in time – and is also an unreasonably high bar. Finally, factor 5 seems to suggest that a claimant can in fact themselves 'put the facts straight' on social media by combatting the defamatory content posted by another – which is arguably not a claimant's responsibility, nor something they may be in the position to do. Indeed, this flies against the very purpose of

[85] *Monroe v Hopkins* (n 22) [71, (2)–(10)].

defamation law itself – that it justifies a restriction to freedom of expression in the face of maintaining individual reputation and, with it, personal dignity. It is hard to see how personal dignity is upheld by the idea that a claimant is expected to wrestle in a virtual mud fight with a defendant online, in a futile attempt to uphold their own (damaged) reputation as a result of a defamatory post.

f. Section 1's introduction in the context of the codified defences in the 2013 Act

It is beyond the scope of this work to consider all of the many changes ushered in by the Defamation Act 2013, most with the aim of bolstering freedom of expression. However, it is important to acknowledge another relevant change that was brought in alongside the introduction of the serious harm threshold: the codification into statute of three long-standing common law defences to defamation. The truth defence is now contained in s 2 Defamation Act 2013, although largely unchanged.[86] What was previously known as the *Reynolds* defence of responsible journalism at common law is now s 4 of the Defamation Act 2013, 'publication on matter of public interest'. The defence of fair comment became 'honest opinion', set out in s 3 of the 2013 Act. In its codification, s 3 dropped the requirement existent at common law that the opinion at issue had to be on a matter of public interest and,[87] unlike fair comment, the new defence is not defeated by malice.[88] The abolition of the public interest requirement provides a not insignificant extension to the types of statements that can be covered by the defence, which was already 'wide' at common

[86] Also see s 5 Defamation Act 1952 (repealed by the 2013 Act).

[87] This does not appear in the (lengthy) text of s 3 as a requirement. On the 'old' position, see: *Joseph v Spiller* [2010] UKSC 53 [3], which refers to *Tse Wai Chun Paul v Albert Cheng* [2001] EMLR 777 [16].

[88] *Joseph v Spiller* (n 87) [4].

law. Indeed, in the recent case of *Dyson*, Mr Justice Kay observed that 'the scope for honest comment, however wounding and unbalanced, was very considerable indeed'.[89] Lowering the bar for defendants attempting to use the defence of honest opinion in tandem with raising the bar for claimants through the introduction of the serious harm threshold results in an undeniable impediment to those wishing in bring an action in defamation, with respect to online content or otherwise. It should, however, be noted that despite this change in s 3, recent cases such as *Butt*, *Dyson* and *Blake* have shown that honest opinion will in many ways be interpreted in a similar fashion to its predecessor, fair comment.[90] Pre-2013 Act case law was extensively used to reach the decision in *Dyson*, while in *Blake* the court reaffirmed the traditional distinction between fact and opinion, finding opinion to be a 'deduction'[91] in a manner consistent with the pre-2013 decision of *Singh*.[92]

The defence that poses a more significant problem for claimants post-2013 codification is s 4. Perhaps unsurprisingly, it is this defence has altered the most from its common law predecessor. At common law, the *Reynolds* defence offered protection for investigative journalists publishing pieces in the public interest that could not be shown to be true (at least to the standard of 'justification' or the truth defence). The defence therefore filled a gap in the law, allowing journalists to publish important stories that the public ought to know, free from the fear of libel. The standard set for reliance on the *Reynolds* defence was that of responsible journalism and it was applied with reference to Lord Nicholls' famous ten-factor list set out in the case of *Reynolds* itself.[93] The

[89] *Dyson v MGN Ltd* [2023] EWHC 3092 (KB) [144].

[90] *Dr Salman Butt v The Secretary of State for the Home Department* [2019] EWCA Civ 933, *Dyson v MGN* (n 89), and *Blake v Fox* (n 6).

[91] *Blake v Fox* (n 6) [23].

[92] *British Chiropractic Association v Singh* [2010] EWCA Civ 350.

[93] *Reynolds v Times Newspapers* (n 83) [205].

factors included: the seriousness of the allegation, its nature, source, whether verification was sought, its status, urgency, whether comment was sought from the claimant, whether the claimant's side of the story was put forth, tone, and circumstances of the publication.[94] *Reynolds* quickly embedded itself as an important defence and generated a range of case law. It is important to note that *Reynolds*, at the most basic level, was about the public's *right to know* certain information and the precise circumstances in which journalists could evade liability in defamation on that basis.[95] The precise scope of the defence was later qualified in a patchwork way by cases that came after *Reynolds*: *Jameel v Wall Street Journal* reaffirmed Lord Nicholls' earlier position that his list of factors should not be treated as an exhaustive set of hurdles, all of which must be overcome to bring a claim.[96] The factors added to the weight of proving responsible journalism had taken place.[97] However, the defence was not without its critics – and in the Libel Reform Campaign, spearheaded by journalists and human rights campaigners as well as scientists, it was argued that the defence was seen as unfairly complex and too difficult to rely on.[98] As a result of the successful lobbying of this campaign, the s 4 defence was born. *Reynolds* and its rich case law was substantively abolished and replaced instead with a two-pronged test in s 4.[99] Firstly, the court must consider whether the information was in the public interest.[100] If the

[94] Ibid factors 1–10 in order.

[95] Jenny Steele, *Tort Law* (1st edn, OUP 2007) 789–90. Credit should also be given to Professor Richard Mullender for elucidating this point in lectures given circa 2013 on the matter at Newcastle Law School, UK.

[96] *Jameel v Wall Street Journal (No 2)* [2006] UKHL 44 [33].

[97] Ibid [32].

[98] House of Lords and House of Commons Joint Committee on the Draft Defamation Bill Session 2010–12, Report (12 October 2011) 8, 22, 27.

[99] Section 4(6) Defamation Act 2013.

[100] Section 4(1)(a) Defamation Act 2013.

answer is affirmative, then the second relevant question is whether the publisher *reasonably believed* that publication was in the public interest.[101]

The replacement of *Reynolds*' rich body of precedent by s 4 has resulted in a number of negative ramifications for the law.[102] Firstly, this author has argued elsewhere that legal clarity for both claimants and defendants has been sacrificed at the altar to appease the Libel Reform Campaign through the enactment of s 4.[103] This is largely because s 4 is too short and it is unclear what the new standard to rely on s 4 *actually is*. Unfortunately, detailed and authoritative guidance from senior courts is still awaited. The Supreme Court in *Serafin* did little to clarify the precise content and scope of the new test in s 4, although Lord Wilson made it clear that s 4 was intended to operate differently and more flexibly to *Reynolds* at common law (but not precisely how).[104] Secondly, reference to responsible journalism in the Act was departed from in order to broaden the scope of the defence and the circumstances in which the defence could be relied upon. Regardless of the merits of this endeavour, this has resulted in complex legal questions that are yet to be answered – it is not yet apparent quite *how broadly* the judiciary will be prepared to interpret s 4. Early indications have shown that s 4 can be applied to a wide range of contextual scenarios, including those where the defendant has no formal (or even informal) journalistic training and is in fact engaging in a

[101] Section 4(1)(b) Defamation Act 2013.

[102] To make matters more confusing, the Explanatory Notes to the 2013 also state that *Reynolds* case law is relevant in an advisory capacity to interpreting s 4: leading some to comment that the defence has only been 'pseudo-codified'. See Fiona Brimblecombe, 'Section 4 Defamation Act 2013: A tale of two approaches' (2024) 29 *Torts Law Journal* 245, Part C, fn 148.

[103] Brimblecombe (n 102).

[104] *Serafin v Malkiewicz* [2020] UKSC 23 [57].

process very far removed from responsible journalism, from which this defence originated. Two prevalent examples where this has occurred are the cases of *Economou* and *Hay*.[105] In *Economou*, de Freitas' daughter had falsely accused Economou of rape and the Crown Prosecution Service were preparing to mount a case against her for perverting the course of justice before she took her own life.[106] On receiving legal advice, de Freitas had supplied information to the press about the matter, including defamatory allegations about Economou, resulting in a number of news articles and broadcasts.[107] Economou brought an action in defamation against de Freitas. The Court of Appeal in the matter allowed de Freitas to successfully rely on the s 4 defence and in so doing extended its scope to encompass what has been dubbed 'contributor immunity'.[108] Further, in the case of *Hay*, it was found that the defendant could rely on the s 4 defence for acting in ways that were merely vaguely reminiscent of an investigative journalist,[109] such as the fact the defendant had undertaken her own informal enquiries and the lack of cooperation of the police in the matter.[110] The case also seems to suggest that because the defendant was writing from her own experiences, that the idea of the information's verification (itself a *Reynolds* factor) should be approached

[105] *Economou v de Frietas* [2016] EWHC 1853 (QB) and *Economou v de Frietas* [2018] EWCA Civ 2591. *Hay v Cresswell* [2023] EWHC 882 (KB).

[106] *Economou v de Frietas* [2018] (n 105) [1]–[9].

[107] Information which de Freitas ardently believed was true. Ibid and see Dominic Garner, 'Case law: *Economou v de Freitas*, Court of Appeal guidance on "public interest" defence' (*Inforrm*, 5 December 2018) https://inforrm.org/2018/12/05/case-law-economou-v-de-freitas-court-of-appeal-guidance-on-public-interest-defence-dominic-garner/ accessed 11 September 2019.

[108] Garner (n 107) and *Economou v de Frietas* [2018] (n 105) [107].

[109] *Hay v Cresswell* (n 105).

[110] Ibid [210, vi].

in a more generous capacity,[111] favourable to a defendant. The turn of s 4 to the expansive is not surprising, as the whole point of jettisoning *Reynolds* at common law and its replacement with s 4 was to extend the scope of the defence and give it a more flexible nature[112] – although the outer limits of the defence as yet remain unclear and unexplained by the judiciary.

It is beyond the scope of this chapter to discuss this matter further and these are just two cases that evidence the potential breadth of the s 4 defence since reform. It is fair to say that the expanding nature of the (now statutory) defences alongside the introduction of the s 1 serious harm threshold has created a problem for claimants in defamation. To add yet more confusion, it appears that s 1 and s 4 at times may yield an interaction: if protection afforded by the s 4 defence in a later 'phase' of publication is found to fall away, it has been held that s 1 may also need reassessment at a later stage of publication. In *Banks v Cadwalladr*, it was found that reliance on s 4 as a defence ought to be reassessed at different points in 'phases' of publication – the defence falling away after new facts came to light which removed a finding of reasonable belief under s 4(1)(b).[113] It was also found in *Banks* that s 1 would also need to be reassessed in light of this, as the contextual circumstances of the later phase of publication had changed.[114] In other words, even if a claimant has successfully argued that s 1 has been met at the outset of a claim (a particularly difficult task for victims of online defamation), if circumstances change in a publication's timeline, then the s 1 threshold may still need to be reassessed.

[111] Ibid [210, xviii].

[112] Lord McNally HL, Grand Committee, vol 741 col 534, and *Serafin* (n 104) [65].

[113] *Banks v Cadwalladr* (n 50).

[114] Ibid [47] and [48].

II. Concluding remarks about s 1 and online defamation

To conclude this section, it is fitting to restate a number of findings from the analysis given about why it may be harder to meet the s 1 threshold in the *defamation by social media* scenario.

1. *Section 1(1) 'raising the bar' has meant that the most effective way to argue that the threshold is met is by the 'evidentiary' route*, providing supporting evidence that shows that the claimant's reputation has been harmed by the post in question. One way to do this is to physically bring people to trial to testify that they thought less of the claimant as a result of the post in question. While this may be hard to do for any claimant, it is particularly challenging in respect of information posted to social media or other websites – as readers may be anonymous, difficult to get hold of, live geographically far away or be less likely to respond to a request to take part in the trial.[115]
2. *It is difficult to know the exact reach of a defamatory post on a publicly available website or social media*, a factor that adds to the weight of s 1 being met – and discussed in three-quarters of judgments.[116] When magazines are sold, this gives a clear idea of how many copies are in the public domain; this is not the case with website 'hits'. Many hits might be from the same small group of people. It may also prove challenging to ascertain where those hits geographically came from, unless IP addresses are tracked – and in any event, people could be using a virtual private network to disguise their location.
3. *Judges have held that proving s 1(1) as satisfied is not merely a 'numbers game', but when online posts are scrutinized for the purposes of the threshold, this often leads to user metrics being*

[115] Sewell (n 4) 60.
[116] Ibid.

adduced – and the precise size of the numbers involved is a significant contributing factor to finding s 1 is met.[117] A reliance on user metrics to this extent is a reductive exercise. Such metrics can also be unreliable as numbers of views do not always correlate to accurately to post readership, especially when posts are viewable in the public domain and not through a logged-in network. Despite this, the courts seem preoccupied with the usage of metrics in online defamation cases to the detriment of other relevant factors. Posts can also be deleted, and with this, valuable information about readership may be lost – which can lead to s 1 failing to be met, a burden which is on the claimant to prove.[118]

4. *Certain cases have found that the s 1 threshold has to be met by each individual statement complained of* rather than holistically if several posts are involved. This fails to take into account that if several successive posts are published around the same time frame, it is likely many of the same readers will have been exposed to all of them – and therefore may form a rounded defamatory impression of the claimant as a result of the *combination* of the posts. The court has also found that s 1 threshold also has to *be re-examined at different publication phases*.[119] This can lead to a finding of s 1 as met falling away at a later stage of proceedings. It should be noted that this is more likely to happen with regards to discussions taking place on social media, where new information comes to light typically more quickly as it can be shared instantly.
5. *There is some indication certain members of the judiciary may harbour a level of inherent scepticism about defamation online*, and information conveyed on social media may be viewed as less serious than information conveyed in a newspaper

[117] *Amersi v Leslie* (n 46) [145] quoting *King v Grundon* (n 62) [40].
[118] Sewell (n 4) 64.
[119] *Banks v Cadwalladr* (n 50) [5].

or other forms of traditional media.[120] When this scepticism is apparent, it can mean that claimant counsel has to work harder to convince a judge that s 1 has been met in the case – but have less avenues to do so as the information has been disseminated through the internet, where it is more challenging to get hard evidence about its impact and reach.

III. The introduction of the single publication rule in s 8 Defamation Act 2013

a. Background to s 8

The single publication rule introduced by the 2013 Act was a truly new addition to libel's legal framework, unlike the serious harm threshold, which was, rather, a stricter interpretation of common law rules that came before it. Section 8 replaces the previous common law position, the multiple publication rule, which termed that each new republication of a defamatory statement by a publisher can yield a fresh claim. By pro-reform campaigners, the multiple publication rule was considered to be unduly skewed towards reputational interests, potentially creating crushing liability on defendants.[121] This is despite the fact that the rule had a clear reason behind its operation; that each subsequent republication, no matter how much later in time, can yield fresh reputational harm – particularly if it is brought to new audiences. The rule itself originated from the famed case of *Brunswick v Harmer*, which concerned a 17-year-old newspaper with libellous content being sold, giving rise to a new claim – with its own limitation period

[120] *Stocker v Stocker* (n 78) [43].

[121] Alastair Mullis and Andrew Scott, 'Tilting at windmills: The Defamation Act 2013' (2014) 77(1) *Modern Law Review* 87, 102.

of one year as generated.[122] The previous position of the law helped to shield individual dignity, the very interest that lies at the heart of reputation's protection in the law. Much like the introduction of the serious harm threshold, this change is yet another example of the 2013 reform's pivot towards Article 10 ECHR interests. The single publication rule in s 8 instead introduces a limitation period of one year, time running from the *first* publication of a defamatory statement by a publisher. If the publisher then later republishes the statement in 'substantially the same'[123] form and an action cannot be brought within this one-year window (from the first publication) then an action is effectively prohibited. Mr Justice Warby states the new rule simply: 'a claim in respect of a defamatory statement is barred once a year has passed since its first publication by the defendant'.[124]

This, of course, may cause problems for individuals looking to action *repetition of statements online over a year later* in the scenario outlined in Chapter 2. Another reason given for single publication rule's introduction was that the law as it stood would freeze the operation of web archives, with the operators of such archives repeatedly exposed to 'fresh' claims in defamation every time information was accessed – however old.[125] The Privy Council explains it as such:

> The effect of section 8 of the 2013 Act in English law is therefore that a claimant has one year from the first

[122] *Duke of Brunswick v Harmer* (n 1); *Loutchansky v Times Newspapers Ltd (No 2)* [2001] EWCA Civ 1805, [2002] QB 783 [57] and Defamation Act 2013. Explanatory Notes, UK Public General Acts, 2013 c 26, Commentary on Sections, Section 8: Single publication rule [60] www.legislation.gov.uk/ukpga/2013/26/notes/division/5/8 accessed 1 December 2024.

[123] Section 8(1)(b) Defamation Act 2013.

[124] *Richardson v Facebook, Google* [2015] EWHC 3154 (QB) [49].

[125] Mullis and Scott (n 121) 102.

> publication of the libel to bring proceedings, however many times the libel is later repeated. This changed the common law position, according to which each publication of a libel gave rise to a separate cause of action with its own one year limitation period …[126]

Section 8 is only short, stating:

> (1) This section applies if a person –
> (a) publishes a statement to the public ("the first publication"), and
> (b) subsequently publishes (whether or not to the public) that statement or a statement which is substantially the same.
> (2) In subsection (1) "publication to the public" includes publication to a section of the public.
> (3) For the purposes of section 4A of the Limitation Act 1980 (time limit for actions for defamation etc) any cause of action against the person for defamation in respect of the subsequent publication is to be treated as having accrued on the date of the first publication.
> (4) This section does not apply in relation to the subsequent publication if the manner of that publication is *materially different* from the manner of the first publication.
> (5) In determining whether the manner of a subsequent publication is materially different from the manner of the first publication, the matters to which the court may have regard include (amongst other matters) –
> (a) the level of prominence that a statement is given;
> (b) the extent of the subsequent publication.[127]

[126] *Hilaire v Chastanet (PC)* [2023] UKPC 22 [53].
[127] Section 8 Defamation Act 2013; emphasis added.

The rule is not based on legal principle, but rather policy. It is a '"legal fiction" crafted to simplify litigation'[128] and reduce it – thereby prioritizing the rights of defendants and freedom of expression. This is despite the fact that online defamation is on the rise and the chance of a statement being repeated online – whether or not it was initially disseminated through the internet by the same publisher – may be high (invoking the *repetition of statements online over a year later* scenario). Interestingly, the new rule mimics a similar rule in American defamation law. This is, of course, a legal system famously swayed in favour of the sacrosanct importance of freedom of speech due to the First Amendment, which does not view reputation as an equal yet competing right. This is unlike the ECHR, which puts both reputation (protected by Article 8) and expression (Article 10) on an equal footing as qualified rights, which is of course the relevant framework for English law. The single publication rule in the US is derived from the common law, but also the Second Restatement of Torts in 1977 and the Uniform Single Publication Act 1952.[129] Much like s 8, the single publication rule in the US works by the limitation period beginning to run 'when a statement is first published'.[130] Harder notes that the US Court of Appeal in New Orleans found in 2007 that the single publication rule also applies to statements published online.[131] Earlier, in 2002, the landmark ruling of *Firth v State* demonstrated a willingness of the American courts to show that the historical, entrenched rules of US defamation law would be applied to online publications as well,[132] similarly holding

[128] Lori A. Wood, 'Cyber-defamation and the single publication rule' (2001) 81(4) *Boston University Law Review* 895, 898.

[129] Ibid 897.

[130] Amy Harder, 'When defamation goes online: Courts are increasingly applying the single publication rule to online publications to deter stale lawsuits' (2008) *The News Media & The Law* 37.

[131] Ibid.

[132] 98 N.Y.2d 365, 775 N.E.2d 463, 747 N.Y.S.2d 69 (2002).

that the single publication rule – with its limitation period of one year – would apply to online posts.[133] In statutory form, the Defamation Act 2013 adopts the more draconian American approach of barring claims after a year has run, regardless of any further reputationally damaging impact on a claimant through successive republication by a publisher and seemingly regardless of the competing interests of a claimant and the defamatory potential of the internet.

b. Thin justifications

As referred to earlier, one of the reasons given for s 8's adoption was to protect those who operated web archives; there was a concern that the pre-existing common law position would curtail archive potential and speech online as a result.[134] Others have likewise argued that without the single publication rule search engines like Google may also be regularly actioned against, which has the potential to dredge up defamatory content from many years ago.[135] This is the famed 'floodgates' argument in tort law, that indeterminate liability of an uncontrolled nature should be restricted at common law on the grounds of policy, rooted in fairness for defendants, avoidance of crushing liability and practicalities. This is also the argument behind the continued use of the rule in the US,[136] Harder observing that 'the purpose of the single-publication rule is to safeguard publishers against endless liability'.[137]

However, it is important to consider if enacting s 8 was the only option to avoid this eventuality or if another way could

[133] Alan J. Pierce, 'New York's appellate courts wrestle with significant issues in internet defamation cases' (2013) 17(4) *Journal of Internet Law* 1, 9.

[134] Mullis and Scott (n 121) 102.

[135] Notwithstanding any operation of the s 5 Defence of Operators of Websites in the Defamation Act 2013 (discussed in Chapter 4).

[136] Pierce (n 133) 12.

[137] Harder (n 130) 37.

have been used. If the reason the single publication rule in the English jurisdiction was (at least in part) based on the desire to maintain the operation of online archives, there were other options that could have been used to safeguard online archivists. Firstly, such operators may have been protected already, if content on the archiving websites had been posted by third parties.[138] If this was not the case – and the content was instead uploaded (even if not authored) by the archivists themselves, then a new defence similar to that in s 5 Defamation Act 2013 could have been drafted to protect online archivists who did not reasonably believe that the documents being archived contained defamatory material; and in the event they were made aware of this, acted appropriately swiftly to remove the information using a notification-and-takedown mechanism. A parallel could easily have been drawn with the rule in *Vizetelly v Mudie's Select Library*,[139] where it was held that even large libraries are expected to take reasonable care regarding republishing defamatory statements of another in books provided in order to be protected from a defamation suit. If this approach was unfavourable, a blanket defence (suitably specific and narrow) regarding online archivists in the pursuance of academic interest and with no malicious intent could have also been drafted. Yet instead, due to the pressure of the reform campaign, parliament chose to impose s 8, which prohibits *all* claims seeking action on republication of a defamatory statement by the same publisher over a year later, thereby impacting reputational interests in a much broader, general fashion. When interviewed by Harder, a US lawyer commented on the court's decision to apply the single publication rule to online posts, stating 'there really wasn't any reason to go a different way'.[140] This ignores the very real threat to individual reputation posed by

[138] By virtue of s 5 Defamation Act 2013.
[139] *Vizetelly v Mudie's Select Library* [1900] 2 QB 170.
[140] Harder (n 130) 37.

the dissemination capabilities of the internet and the ease and scale by which information can be republished to the masses online, damaging individual reputations repeatedly. Further, when one examines justifications for operating the rule in a US context, it is plain to see that many such justifications do not apply to the English jurisdiction. Wood has observed that the American rule is concerned with

> preventing a plaintiff from bringing numerous suits concerning the same publication. Instead, the rule folds the claims from all applicable jurisdictions into one claim. In addition, the rule precludes the recovery of excessive damages, a possible result of numerous suits resources by preventing depletion through duplicative actions.[141]

The multiple jurisdiction issue applies to the US as different suits can be opened in different states, an issue which is not applicable to the English and Welsh jurisdiction. Wood's second observation – that the rule in part bars crushing liability through the recovery of excessive damages – also fails to translate to the English legal system, where monetary damages are typically far lower than in the US. As these justifications fall away, and the online archive issue could have been dealt with by other legal means, it is hard to understand why it was necessary to enact such a draconian shift to the law with s 8 when online libel is more prevalent than ever before. The central problem with the single publication rule in s 8 is that it tips too far in favour of expression, with the potential to bar redress and the protection of Article 8 interests. Phillipson has called for an 'equitable' reinterpretation of s 8 by the courts, using s 3 of the Human Rights Act 1998 to give

[141] Wood (n 128) 898.

credence to Article 8 concerns.[142] At the time of writing, this unfortunately has not come to fruition, and seems unlikely to do so.

c. What is republication in 'substantially the same' form?

If one examines the text of s 8, there is a clear caveat to its operation. It only applies to defamatory statements as (re)published if they are 'substantially the same'.[143] It goes on to stress that the section is not applicable if the second publication is 'materially different' from the first publication.[144] It has not been clarified to date by the courts what publishing something in substantially the same form actually looks like, nor how different a publication must be to be considered materially different. Under the US single publication rule, it would require 'a substantial modification of the same version',[145] as 'merely adding nonsubstantive material' would not suffice.[146] This is an extremely high bar for claimants to meet, but there is no evidence as yet that this will be the English position adopted under s 8 – and it is submitted here it should not be. A sympathetic reading of the extent of difference required under s 8 to disapply the limitation period may be necessary in order that a claimant can seek redress under the *repetition of statements online over a year later* scenario. This would at the very least go some way

[142] Gavin Phillipson, 'The "global pariah", the Defamation Bill and the Human Rights Act' (2012) 63(1) *Northern Ireland Legal Quarterly* 149, 182.
[143] Section 8(1)(b) Defamation Act 2013.
[144] Section 8(4) Defamation Act 2013.
[145] Notes, 'The single publication rule and online copyright: Tensions between broadcast, licensing, and defamation law' (2010) 123(5) *Harvard Law Review* 1315, 1317, and see *Atkinson v McLaughlin*, 462 F Supp 2 d 1038, 1052 (DND 2006); *Nichols v Moore*, 334 F Supp 2d 944, 952 (ED Mich 2004).
[146] Notes (n 145) 1318.

to readdress the balance between the rights of defendants and ability of claimants to bring a suit in light of the limitless defamatory potential of the internet.

Phillipson has argued that republishing something on the internet when it has previously been published in traditional media should constitute a material difference for the purposes of negating s 8, as placing something on the internet has the potential to do new and significant reputational damage, although this was not a view shared by the Joint Committee on the (then) Defamation Bill.[147] Kumar agrees with Phillipson's proposition, on the basis that the statement, when republished online, will reach a 'new audience'.[148] Aside from this, the manner of publication moving from (for example) paper print to the internet is clearly also physically different from a pragmatic perspective. Perhaps surprisingly, this view is consistent with the US approach to the single publication rule, as this qualifies as a 'republication exemption',[149] as putting something online 'reaches a new group'[150] and is therefore classed as a new and separate publication, as it is 'the result of a conscious independent act'.[151] While this is logically and principally is the legally correct approach, any contrary finding with regards to s 8 (as suggested by the Defamation Bill's Joint Committee)[152] would in fact find English defamation law is going *further* to shield defendants and disadvantage claimants than US law, despite the operation of the ECHR in the English jurisdiction.

[147] Phillipson (n 142) 182.

[148] Sapna Kumar, 'Comment: Website libel and the single publication rule' (2003) 70 *University of Chicago Law Review* 639, 657–9 and Notes (n 145) 1319.

[149] Wood (n 128) 901.

[150] Ibid.

[151] Notes (n 145) 1318 and *Lehman v Discovery Commc'ns, Inc*, 332 F Supp 2d 534, 539 (EDNY 2004).

[152] Phillipson (n 142) 182.

Much like the s 1 threshold, the introduction of s 8's limitation period could not have come at a worse time for those seeking to litigate about reputational damage caused by an online publication. Defamation can now happen instantaneously, easily and frequently on the internet – yet a claimant may be barred from bringing an action against a republication of an online post if the court views the statement as substantially the same. In order to mitigate disadvantages to claimants in this scenario, it is argued here that courts must interpret 'substantially the same' under Article 8(1)(b) to include exact copies of the statement, but not posts that have been otherwise changed in terms of content and format. This could include something published in a shorter, abridged version or presented in a different way – s 8(5)(a) suggests itself that a change to presentation could warrant a finding that s 8 is disapplied.[153] It is argued here that the bar for what constitutes a change in presentation should be not set too highly. This liberal interpretation would be justified by the fresh reputational harm adduced by the subsequent republication – reaching a different online audience and damaging personal dignity afresh, when the initial publication may have been long forgotten. For these reasons, it is also argued that republication by the same publisher of the statement online should also warrant a finding that this is materially different, and s 8 disapplied. There may be some prima facie support for this approach in s 8(5)(b) itself, which notes that 'extent' of publication is relevant to establishing whether the publication is materially different.[154]

d. A mitigating factor: s 32A of the Limitation Act 1980

There is one mitigating factor that may work to ease concerned claimants in a *republication online over a year later* scenario, albeit

[153] Defamation Act 2013.
[154] Ibid.

in a narrow category of cases. In *Denman*, Sir David Eady observed that there is still a way to avoid the application of s 8's limitation period, even if the statement that is republished is substantially the same. Sir Eady noted that by using s 32A Limitation Act 1980 a judge can hold that despite s 8's limitation period, it is still 'equitable' for the case to proceed.[155] However, this is entirely at the discretion of the court and for this workaround to apply circumstances have to be of an 'exceptional nature'.[156] In *Denman*, the court observed that the leave obtained under s 32A is only granted in a narrow set of circumstances: 'The court must have regard to any prejudice that would be caused to either side. It has been long recognised that the exercise of that discretionary jurisdiction is exceptional because, for reasons of public policy, time is treated as being of the essence in the defamation context.'[157]

In other words, judges will be slow to grant leave under s 32A of the Limitation Act 1980, unless there is a compelling or unusual reason why there has been a delay in bringing proceedings. The court will not be persuaded by those who want to 're-fight old battles'.[158] From the text of 32A itself, it is clear that the court is expected to undertake a delicate balancing exercise between the rights of claimants and defendants and assess to what extent it will cause them prejudice in disapplying the limitation period.[159] In *Hemming v Poulton*, the High Court stated that 'it is recognised that a court should be hesitant to exercise its discretion under s.32A'.[160] This is because the purpose of a libel action, according to the court,

[155] *Denman v Associated Newspapers Ltd* [2016] EWHC 2819 (QB) [8].
[156] Ibid.
[157] Ibid [7]. Also see: *Steedman v BBC* [2001] EWCA Civ 1534, *Austin v Newcastle Chronicle and Journal Ltd* [2001] EWCA Civ 834, *Bewry v Reed Elsevier* [2015] 1 WLR 2565 [5]–[8].
[158] *Denman v Associated Newspapers* (n 155) [10].
[159] Section 32A (1)(a) and (b) Limitation Act 1980.
[160] *Hemming v Poulton* [2023] EWHC 3001 (KB) [131].

is to achieve swift justice.[161] It is hard to see how this reason would be served by barring a claim (which would deliver no justice whatsoever). More charitably, courts are in the difficult position of considering allowing a needy claimant's action to proceed, while at the same time attempting not to undermine the will of parliament with s 8. The final point, then, is that this exception to the rule of s 8 will only be applicable in a narrow set of circumstances. This may not give much cause for hope to claimants with respect to a *republication online over a year later* scenario, as the balancing exercise in s 32A is a difficult one to meet and looks to only disapply s 8 in rare eventualities.

IV. Concluding remarks for Part I

Part I of this chapter has focused on two developments in libel reform introduced by the Defamation Act 2013: the introduction of the serious harm threshold in s 1 of the Act and the single publication rule in s 8. The serious harm threshold has quite significantly raised the bar for claimants generally, with the evidentiary route for proving the threshold has been met still the most likely way s 1 can be satisfied. Claimants litigating in a *defamation by social media* scenario may find the threshold particularly difficult to meet, due to these evidentiary requirements and the struggle of otherwise making out an inferential case. The courts' preoccupation with engagement statistics on social media has not served to help matters as data is not always reliable and can obfuscate more pertinent issues. Some members of the judiciary also seem to have a tacit mistrust of social media communication, with the preordained belief that statements posted on social media may be inherently less serious; once again making meeting this threshold harder. In the *republication online over a year later* scenario, claimants are also met with serious hindrances to their claim. Unless they can establish

[161] Ibid.

that the second defamatory publication in question is materially different from the first publication by the same publisher, then the claim will be barred as it is outside s 8's statutory limitation period of one year. It is as yet unclear exactly how different two publications must be in order not to be deemed substantially the same; more guidance from an authoritative court is awaited. Raising the threshold claimants must meet to bring an action, and limiting the time they have to do it, are both draconian ways to reduce the ability of claimants to bring actions in defamation, be these in respect to statements posted on or offline. Both of these developments have created a 'perfect storm' for claimants seeking redress for online defamation. It is now harder than ever to bring an action, at a time when the unbridled dissemination abilities of the internet can communicate defamatory content worldwide anytime, anywhere and at a click of a button.

★★★

Part II: Liability of host websites and defamation by an AI tool

Part I of this chapter focused on two *legal* developments in the Defamation Act 2013 reform that have made bringing an action in online defamation more challenging. Part II will consider two *technological* phenomena of online defamation that may be present in an action. The first instance that will be considered is when a defamatory post is published by a third party onto a 'host' website – can or should action be taken against that website? Many websites now have capabilities for users to leave comments under posts or articles and other capabilities for users to publicly interact with one another, such as virtual worlds. Newspaper websites, such as the *Mail Online*, often have active comment sections.[162] Responding to the posts of

[162] www.dailymail.co.uk/home/index.html accessed 1 December 2024. At the time of writing, the *Mail Online* had 22.4 million 'followers' on Facebook.

others is a key functionality of social media websites such as X, while virtual worlds can now be entered and used as conduits to disseminate defamatory information.[163] However, it may not always be possible or desirable to bring an action against a third party who has posted a comment. It will be discussed here whether a claimant in the *third-party poster* scenario can bring an action instead against the website operator (or host) themselves. The second phenomenon that will be considered is the burgeoning ability for individuals to be defamed by an AI program. As discussed in Chapter 1, the technological capabilities of the internet have advanced at a staggering pace. It is now entirely possible that a program powered by AI can communicate defamatory statements about another person, either directly to a user or posted to the internet more widely. AI regularly 'hallucinates' and creates false information about living people when responding to text prompts. This section will consider who should be actioned against by a claimant in such an eventuality. This is the *defamation by AI tool* scenario. Both of these technological issues as present in an action will have a pivotal impact on the success of a claim and how it is assessed.

I. The defence for operators of websites under s 5 Defamation Act 2013

a. A new defence

In the *third-party poster* scenario, claimants are seeking redress against a defamatory statement that has been posted to a website by a third party who is potentially anonymous or difficult to contact. Section 5 of the Defamation Act 2013 gives a new defence to website operators in these circumstances, in the

[163] See, for example, the Metaverse: Meta, 'What is the Metaverse?' https://about.meta.com/uk/what-is-the-metaverse/ accessed 1 December 2024.

event that they may be claimed against instead. The defence is applicable if the operators did not post that defamatory material themselves and complied with the annexed statutory regulations.[164] Section 5 states:

> (2) It is a defence for the operator to show that it was not the operator who posted the statement on the website.[165]
> (3) The defence is defeated if the claimant shows that ...
> (b) the claimant gave the operator a notice of complaint in relation to the statement, and
> (c) the operator failed to respond to the notice of complaint in accordance with any provision contained in regulations.[166]

In other words, in order for a website operator to rely on the defence against an action from a claimant, in the *third-party poster* scenario they must show that they followed the procedure outlined by the statutory instrument annexed to s 5.[167] The statutory instrument observes that with 48 hours of a notice from a prospective claimant concerning (potentially) defamatory content posted on the website,[168] the web host must send notice of that complaint to the third-party poster.[169] An explanation should also be sent to the poster that the statement will be removed from the website,[170] unless there is a response in writing to protesting the contrary by the end of

[164] The Defamation (Operators of Websites) Regulations 2013, Statutory Instrument No 3028 www.legislation.gov.uk/uksi/2013/3028/contents/made accessed 1 December 2024.
[165] Sections 5(2) Defamation Act 2013.
[166] Sections 5(3)(b) and (c) of the Defamation Act 2013.
[167] The Defamation (Operators of Websites) Regulations (n 164).
[168] Ibid Schedule 2(1).
[169] Ibid Schedule 2(1)(a).
[170] Ibid Schedule 2(1)(b).

the fifth day after the notification is sent.[171] If the poster does not wish for the content to be removed they must provide their full name and address,[172] so that they can be actioned against as an individual in defamation by the prospective claimant. If the third-party poster fails to respond by the end of the fifth day, 48 hours later the statement complained of must be removed from the website by the host.[173]

It is argued that this is one of the few examples of good practice in the 2013 Act, which seeks to balance the interests of both those who have been defamed online on the one hand and, on the other, the expression rights and economic interests of host websites and users. It allows for a potentially swift remedy for claimants in the form of the information's removal from a website at the end of the fifth day, absent of any contrary protestations from the poster (and the provision of their contact details). In the event a third-party poster objects to the removal of the statement, they need only request this and provide their details, such that an action can be brought against them individually in defamation – and have their day in court. This means that the statement remains on the website in this interim period. Interestingly, much like s 8 of the Defamation Act 2013 and the US single publication rule, there is a counterpart US law that mirrors the s 5 defence. Section 230 of the Communications Decency Act 1996 in America states that 'no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider'.[174]

---

[171] Ibid Schedule 2(1)(b)(i).

[172] Ibid Schedule 2(2)(b)(i) and (ii).

[173] Ibid Schedule 5(1) and (2)(a).

[174] Communications Decency Act of 1996, 47 USC § 230 (2016) and Dallin Albright, 'Do androids defame with actual malice? Libel in the world of automated journalism' (2023) *75 Federal Communications Law Journal* 103, 108.

b. Potential issues

Section 5 largely succeeds in treading a narrow line between competing rights, but it too has imperfections. Given the unlimited defamatory potential of the internet and the rise in defamation on social media and other websites, it was unwise of the section to adopt the default position that if a statement's removal is contested, it remains visible on the host website while a claimant brings an action against the third-party poster. This does nothing to readdress the clear imbalance in favour of Article 10 ECHR rights in the digital age, or to lessen the effect of the 2013 Act reform that prioritizes expression at the expense of reputation. With every day the potentially defamatory statement remains online, the more harm to personal dignity and one's self-perception can be rendered,[175] as the statement may continue to reach new audiences – or could even be subject to an upwards trend and 'go viral' and reach hundreds of thousands of new readers.[176] This position is also out of step with the privacy tort of misuse of private information in England and Wales. In actions for misuse of private information, a claimant can apply to a court to have an interim injunction that prohibits disclosure of the complained-of material pending trial.[177] Although these applications are at the discretion of the court (and are not always successful), long-standing common law rules show that obtaining a comparable interim injunction in a defamation case is not possible. This divergence of approach

[175] See Chapter 2.

[176] Reaching tens of thousands of readers in a short space of time is considered 'going viral' on X. See 'How many views is viral?' (*Fourthwall*, 6 March 2024) https://fourthwall.com/blog/how-many-views-is-viral accessed 1 December 2024.

[177] For the current test for such an order, see *Cream Holdings Ltd v Banerjee and Others* [2005] 1 AC 253 [22] and *NPV v QEL and ZED* [2018] EWHC 703 (QB).

is further compounded by s 5's position that complained-of posts must remain visible, pending trial – which once again disadvantages personality interests.

At just over a decade old, the Defamation Act 2013 is still considered 'new' in the context of English defamation law. As such, academics and practitioners alike await guidance about many of the new defence's parameters. One issue that remains unclear is if the defence will be considered applicable to large social media websites such as Facebook, X, YouTube, Instagram and virtual online worlds such as the Metaverse. The court in *Goldsmith v Bissett Powell* observed 'there is no authority that I am aware of relating to Facebook administrators and whether they fall within s 5'.[178] In a *third-party poster* scenario, a claimant may find that a defamatory post has been published to a public network on Facebook or the Metaverse, but the poster themselves is not directly contactable to serve an action against. They may operate their posting account under a pseudonym, giving out no other details on the given social media site, and remain unresponsive to direct messages from the person defamed. If this is the case, it is argued here that this defence *should* be applicable to a large social media website or virtual world host. This would incentivize a large platform such as Facebook or its parent company Meta to directly contact a poster as per the statutory instrument on receipt of a complaint, which will more likely engender a response due to the perceived power imbalance between a poster and a large conglomerate. Ultimately, this will lead to the defamatory statement's removal or sufficient contact details of the third-party poster being passed on to the claimant so an action can be brought directly. This at the very least is a clear route to some potential redress in court on the part of a claimant.

[178] *Goldsmith v Bissett-Powell* (n 52) [169].

c. Approach of the Strasbourg Court

To date, there have been two high-profile decisions at the European Court of Human Rights (ECtHR) which consider host website liability for third-party comments posted to online news portals: *Delfi* and *Magyar*.[179] The point that arises from both Strasbourg cases, at least implicitly, is that online content providers or 'hosts' can be subject to civil liability for infringement of personality rights through comments posted by others.[180] These provide an interesting point of comparison for the s 5 defence for web hosts, in terms of how far the defence extends and what sort of responsibilities the English courts in future may impose on website providers in relation to defamatory material posted by others to their sites. As yet there is little case law on the s 5 defence, but it seems likely in future that the precise scope of the defence will be tested in courts, perhaps by large web operators who wish to argue the defence is applicable. There is much academics and practitioners do not know about the s 5 defence, which must be dictated by the passage of the common law. In *Delfi*, Strasbourg's Grand Chamber upheld the earlier First Section's judgment that the Estonian national courts' decision to hold an internet news portal liable for offensive comments made under an article by third parties was *not* a violation of Article 10 ECHR.[181] In *Magyar*, a year later, the opposite finding was reached: the Hungarian courts, in holding a regulatory body and a news portal liable for third-party defamatory comments posted by others to the news website, *were* in violation of the right to freedom of expression.[182] Although at first glance

[179] *Delfi AS v Estonia* App no 64569/09 (16 June 2015). *Magyar Tartalomszolgaltatok Egyesulete v Hungary* App no 22947/13 (2 February 2016).

[180] Also see Andras Koltay, *New Media and Freedom of Expression* (Hart 2019) ch 6.

[181] *Delfi AS v Estonia* (n 179) [162].

[182] *Magyar v Hungary* (n 179) [91ff].

these decisions may seem contradictory, a closer examination reveals this is not the case. The deciding factors of finding no violation (in *Delfi*'s case) and the contrary finding in *Magyar* were rooted in the degree of balancing domestic courts had conducted with a view to assessing the competing Article 8 rights of those defamed or threatened by the comments on the one hand, and the rights of internet service providers, website hosts and internet users more broadly under Article 10 on the other. Perhaps the most important aspect of both decisions are the balancing factors the ECtHR expounded in *Delfi* – and reused in *Magyar* – in order to guide Article 8 and 10 balancing in such a situation. These factors could provide a framework for English courts to employ (as at the very least, a starting point) when adjudicating future decisions on the scope or applicability of the s 5 defence.

By way of background, *Delfi v Estonia* was handed down by the Grand Chamber in 2015,[183] on appeal from the First Section decision in 2013.[184] The matter concerned an online news website named Delfi, a prolific Eastern European news site with a wide readership.[185] It was straightforward for third parties to add comments under a news article on Delfi's platform. People could choose their own name or a pseudonym, the provision of an email address was not essential and comments were uploaded by default.[186] As such, comparatively little oversight by Delfi was necessary for comments to appear under news stories published on the website. The few safeguards Delfi did have in place were that certain obscene words were filtered and there was a notice-and-takedown mechanism for 'hate' comments,[187] encompassing comments or words of the most

[183] *Delfi AS v Estonia* [2015] (n 179).
[184] *Delfi AS v Estonia* App no 64569/09 (ECHR, 10 October 2013).
[185] Ibid [7].
[186] Ibid [8].
[187] Ibid [9].

offensive kind.[188] The matter at issue concerned an article Delfi had posted about a shipping company which it claimed had quashed plans for an ice road in Estonia.[189] Twenty third-party comments posted under the article personally attacked person 'L', a significant shareholder at the company,[190] using threatening and abusive language.[191] The national courts in Estonia found that Delfi was a publisher of the third-party posts and damages were awarded.[192] Importantly, the Grand Chamber decision in *Delfi* upheld a number of balancing factors articulated by the earlier 2013 decision in finding Article 10 had not been infringed.[193] The court went on to apply these factors to the matters at hand.

Firstly, the Grand Chamber considered the factor of the *context of the comments*. Under this headline, the court deemed it relevant that Delfi had control over the comment section on its website and was therefore in a position to moderate content uploaded by others.[194] It was deemed important that Delfi was a popular website that often attracted many comments, and that more comments meant more website traffic and therefore more advertising revenue for Delfi's benefit.[195] Delfi as a website was involved in content moderation, had regulations for third-party comments and did not play the role of a merely passive intermediary.[196] Secondly, the Grand Chamber considered the factor of *liability for authors of the comments*. Under this factor, it was seen relevant that it was difficult to bring an action against

[188] Although these filters were later deemed ineffective in *Delfi AS v Estonia* [2015] (n 179) [156].
[189] *Delfi AS v Estonia* [2013] (n 184) [12].
[190] Ibid [14].
[191] Ibid [13].
[192] *Delfi AS v Estonia* [2015] [26].
[193] Ibid [64] and *Delfi AS v Estonia* [2013] (n 184) [86].
[194] *Delfi AS v Estonia* [2015] [144].
[195] Ibid.
[196] Ibid.

the authors of the comments themselves due to the anonymous nature of the internet.[197] Therefore, actioning against Delfi was seen as a legitimate and necessary alternative. The third and perhaps most important factor applied was *measures taken by applicant company*. The thread in question would have likely drawn the attention of Delfi staff, as the comment section was unusually active.[198] Despite this, there was still a delay in Delfi removing the offending comments. The word filter employed as a method of offensive content moderation was regarded by the court as unfit for purpose, as offensive comments had slipped through the net, as was the notice-and-takedown policy – which the court viewed as too lengthy a process to sufficiently safeguard against hateful comments such as those at issue.[199] Finally, following the First Section, the Grand Chamber considered the factor of *consequences for the applicant company*. They found that there were no significant long-term negative effects for Delfi as a result of the Estonian courts' decision, as a fine was issued but Delfi did not in fact need to change its business model in order to accommodate the ruling. Damages for non-pecuniary loss had not been awarded and, as such, the remedies enforced were far from draconian.[200] An important point to emphasize regarding *Delfi* is that the comments complained of amounted to hate speech,[201] and were 'low-value' speech,[202] and for this reason, the decision was not as controversial as some have attempted to claim.[203] While some defamatory statements may tread the line of hate

[197] Ibid [147].
[198] Ibid [152]–[153].
[199] Ibid [156]–[159].
[200] Ibid [160].
[201] Ibid [159].
[202] In the form of insults and threats.
[203] Lorna Woods, 'The *Delfi AS v Estonia* judgment explained' (*LSE Media Policy Project Blog*, 16 June 2015) https://blogs.lse.ac.uk/mediapolicyproject/2015/06/16/the-delfi-as-vs-estonia-judgement-explained/ accessed 2 September 2019, and Neville Cox, '*Delfi AS v Estonia*: The liability of

speech, many reputationally damaging statements that are actionable under English law would not. The result of the hateful nature of the comments in *Delfi* was that the failure to remove the comments quickly was seen by the court as particularly egregious. As a result of the application of these factors, the court concurred with the First Section's 2013 finding that Article 10 had not been violated.

This method of analysis was followed in the decision of *Magyar* a year later. Here, the court found that the Hungarian national courts had violated freedom of expression by finding a regulatory body of internet service providers and a news portal liable for third-party comments, due to errors in balancing rights conducted by the Hungarian courts.[204] In finding whether an appropriate balance had been struck between Article 8 and Article 10, the Strasbourg Court utilized each of the balancing factors listed by *Delfi* in order to guide their analysis.[205] In *Magyar* the court found that the article which generated the third-party comments in question was on an important matter of public interest (factor: *context in which the comments were posted*).[206] The court also found that it made sense in this case to find those who had made the comments themselves legally responsible according to the *liability for authors of the comments* criterion.[207] Additionally, the website in *Magyar* had removed the comments as soon as it was told about them (*measures taken by applicant website*),[208] therefore acting more swiftly than Delfi. In terms of the *consequences* factor, the court observed in *Magyar* that the claimant's reputation had already

secondary internet publishers for violation of reputational rights under the European Convention on Human Rights' (2014) 77(4) *Modern Law Review* 619.

[204] *Magyar v Hungary* (n 179) [2].

[205] Ibid [63]–[71].

[206] Ibid [72].

[207] Ibid [78].

[208] Ibid [80].

been significantly damaged in light of other comments not complained of and that a threshold of seriousness had not been applied by domestic courts when examining the case.[209] In all senses, the facts of the case in *Magyar* were weaker according to the balancing factors articulated by the court than those in *Delfi*. The judgment in *Magyar* is a demonstration of the breadth and flexibility of the *Delfi* factors and how they can be applied to different facts to reach diametrically opposed conclusions. *Magyar* is also useful as a further example of application of the *Delfi* factors in the event that English courts seek to rely on them for reference when adjudicating a future s 5 dispute. The *consequences* of the decision appear to be particularly important from the perspective of the ECtHR (for both claimants and defendants), as do the *nature and severity* of the comments and the likelihood of it being *realistically possible* to bring an action against third-party posters themselves. Finally, it should be noted that the ECtHR (somewhat controversially) went yet further in the case of *Sanchez v France* in 2023, when the Grand Chamber found that Article 10 ECHR had *not* been violated by the French courts imposing a criminal sanction on Mr Sanchez, a local councillor, after he did not remove hate speech from his Facebook page despite it being posted there by third parties.[210] In the case, the relevant French laws were found to purse the legitimate aims of the reputation of others, as well as the prevention of crime.[211] This decision may, however, be explained with reference to the particular facts of the case. *Sanchez* concerned hateful Islamophobic comments posted to a Mr Sanchez's Facebook wall during election time, when Mr Sanchez himself – already involved in local politics – was running for parliamentary election.[212]

[209] Ibid [85].
[210] *Sanchez v France* App 45581/15 (15 May 2023) [3].
[211] Ibid [144].
[212] Ibid [11] and [14ff].

The crime in France he was convicted for was the incitement of violence or hatred on the grounds of religion.[213] The facts of the case are more aligned to hate speech rather than falling clearly into the remit of defamation law, perhaps even more so than in the case of *Delfi* itself (where certain comments were also found to be threatening and hateful, therefore also crossing the line into hate speech).[214] In making its finding of no violation, the ECtHR noted it relevant that Mr Sanchez was involved in politics,[215] the court finding that for politicians, 'when expressing themselves in public, to avoid comments that might foster intolerance'[216] and to 'foster the exclusion of foreigners constitutes a fundamental attack on individual rights, and everyone – politicians included – should exercise particular caution in discussing such matters'.[217] In the decision, *Delfi*'s precedent was considered and the factors discussed there were applied.[218] *Context of the comments* was considered (including their nature),[219] the *political context* was also considered as well as *specific applicant liability for third-party comments*.[220] *Steps taken* by Sanchez were also examined,[221] as well as the *possibility of holding the authors liable*,[222] and the *consequences* of the domestic decision.[223] The *Delfi* criteria for host liability is therefore still going strong and has shown through *Sanchez* that it can be applied to render liability for *private individuals* who can also seen as web hosts (in this case, Mr Sanchez and his Facebook

[213] Ibid [3].
[214] Ibid: Part B of the Grand Chamber's decision that discusses hate speech specifically [60ff].
[215] *Sanchez v France* (n 210) [149]–[150].
[216] Ibid [150]
[217] Ibid.
[218] Ibid [169ff].
[219] Ibid [169].
[220] Ibid [179].
[221] Ibid [190].
[222] Ibid [202].
[223] Ibid [205].

wall), as well as large conglomerates. It is important, however, to note that the facts of *Sanchez*, given the political context and nature of the comments, were particularly egregious.

Interestingly, the Grand Chamber's decision in *Delfi* differentiates online news portals from social media websites on the basis that social media sites do not generate their own content.[224] It is respectfully submitted here that this distinction is outdated and out of touch with the practical operation of social media websites today. Many social media sites mix user-generated posts and posts generated by the host platform, either by company employees themselves or by outsourcing to freelance content creators. An example of such content would be 'YouTube Rewind', which is the tradition of YouTube itself releasing a video recapping memorable yearly events through a YouTube narrative, largely by employing the talents of independent creators on the site to star in the video itself.[225] Similarly, Facebook now embeds articles from other websites in its landing page and displays advertisements.[226] With the introduction of rapidly evolving AI programs across the internet, the distinction between social media, virtual worlds, news portals and other types of website on which users can leave comments is set to become increasingly blurred. For these reasons, it is argued here that as helpful as the laundry list of factors laid out in *Delfi* (and followed by *Magyar*) may be, they provide far from all the answers to disputes that are yet to come about the scope and application of the s 5 defence. In any event, for these reasons, alongside those articulated earlier, it is argued here that the s 5 defence should be operable not only with regard to the likes of news portals but also social media websites such as

[224] *Delfi AS v Estonia* [2015] [112] and [116].

[225] See, for example, 'YouTube Rewind 2018: Everyone controls Rewind' www.youtube.com/watch?v=YbJOTdZBX1g accessed 9 January 2025.

[226] Facebook, 'Branded content, "Overview"' www.facebook.com/facebookmedia/get-started/branded-content accessed 23 April 2018.

X, YouTube, Facebook, Instagram and, by extension, virtual worlds (which operate as wider-scale social media sites) where users can interact and disseminate defamatory information. Applicability of the s 5 defence to these types of sites provides a clear incentive for large website operators to act swiftly to provide third-party poster contact information so that they can be individually actioned against in defamation – or else remove the offending post in question. Helpfully, the more recent case of *Sanchez* at the Grand Chamber has in fact shown the willingness of the Strasbourg Court to extend *Delfi* principles to social media websites (Facebook), to at least render private-individual 'hosts' liable for third-party comments.[227]

II. Defamation by an AI tool

a. A rising threat

While AI begins to demonstrate the staggering scale of its capabilities throughout the world,[228] many questions about online libel remain unanswered. The challenge of identifying an appropriate defendant to bring an action against in the *defamation by AI tool* scenario is complex. It seems unclear as yet whether existing legal frameworks in England and Wales can effectively tackle such an eventuality. It should be noted at this stage that this book is primarily concerned with generative AI and its defamatory capabilities. There are undoubtedly pressing problems with analytical AI tools and

[227] *Sanchez* (n 210).

[228] For example, on the day of writing, OpenAI made a video generator publicly accessible. See Dara Kerr, 'OpenAI makes AI video generator Sora publicly available in US' *The Guardian* (9 December 2024) www.theguardian.com/technology/2024/dec/09/openai-ai-video-generator-sora-publicly-available accessed 10 December 2024.

their concerning tendency towards bias and discrimination[229] – but this is outside the argument advanced in this work. The rapid advancement of AI and its acceptance in everyday life shows no sign of halting. Respected broadsheet newspapers, such as *The Guardian* and, in the US, *The Washington Post*, have already trialled AI-written articles.[230] However, it is obvious that without rigorous oversight, fully automated production of AI news pieces could lead to false information presented as fact.[231] As AI programs are primarily used, downloaded and accessed through the internet, the public's increased reliance on the internet compounds the implications for personal dignity when one is defamed online in the *defamation by AI tool* scenario.[232] As noted in Chapter 1, the attention span of the 'digital generation' has reduced – and spreading disinformation online may be particularly effective in part due to short durations of engagement,[233] where individuals accessing the defamatory content do not have time to question its accuracy. Volokh observes that companies which create AI programs regularly promote their programs as reliable in order to ensure they are marketable; it therefore follows that this invites legal responsibility when users believe untrue, defamatory claims produced by that AI system.[234] In light of this, it is not a valid defence for companies that create AI programs to then argue later at trial that information was never intended to be completely accurate, or held out to be so.[235]

[229] See, for example, Nima Kordzadeh and Maryam Ghasemaghaei, 'Algorithmic bias: Review, synthesis, and future research directions' (2022) 31(3) *European Journal of Information Systems* 388.
[230] Albright (n 174) 104.
[231] Ibid 109.
[232] Ibid 110.
[233] Ibid.
[234] Eugene Volokh, 'Large libel models? Liability for AI output' (2023) 3 *Journal of Free Speech Law* 489, 498.
[235] Ibid 499.

Indeed, English defamation law would not require everyone who read an AI program's defamatory statement to believe it to be true; rather, it is instead a test of what the 'reasonable reader' believed.[236] The defamatory potential of generative AI systems is also concerning because of their power to sway thought through their large outreach online.[237] As Albright observes, 'dissemination follows naturally' on social media.[238]

There are many different types of AI system, with varying levels of human involvement. So-called semi-autonomous production allows the AI system room to generate its own content more freely than other types of production and to publish information without oversight. As a result, this type of AI program can lead to an increased likelihood of defamatory content being generated. This type of program is also seen as particularly valuable in industry, as it has the potential to generate content that feels more 'natural', 'human' and ultimately convincing – with less costs for overheads before publishing.[239] The most popular type of AI program currently is 'a pretrained generative model on the whole internet', where input data is 'only lightly curated'.[240] Examples of such systems are predictive text models like ChatGPT-2 and 3.[241] Predictive text models often do not 'opt-out' of answering user-prompted questions and instead may fabricate information in order to (one can only assume) satisfy the end-user.[242] Language models

[236] Ibid.

[237] Inyoung Cheong, Aylin Caliskan and Tadayoshi Kohno, 'Safeguarding human values: Rethinking US law for generative AI's societal impacts' (2024) 5 *AI and Ethics* 1433 https://doi.org/10.1007/s43681-024-00451-4 accessed 10 December 2024.

[238] Albright (n 174) 110.

[239] Ibid 108.

[240] Peter Henderson, Tatsunori Hashimoto and Mark Lemley, 'Where's the liability in harmful AI speech?' (2023) 3 *Journal of Free Speech Law* 589, 620–2, 602, and Volokh (n 234).

[241] Ibid Henderson (n 240).

[242] Ibid 603.

can also be influenced by training websites that contain untrue, hateful or offensive speech that companies have failed to weed out at an earlier stage.[243] As Cheong, Caliskan and Kohno have observed, error testing is expensive, lengthy and difficult to resource, particularly for small companies.[244]

b. The nature of the threat

The prevailing facet of generative AI that is leading it to publish potentially defamatory content is the phenomena of AI 'hallucinating'. In other words, AI is regularly observed creating false information in response to prompts – including that which concerns other people – and cannot be traced back to input data.[245] It is as yet unclear how or why AI programs do this, although it is a disturbingly common occurrence. This false material is then presented as fact, often confidently, to the AI tool user – and it is impossible prima facie to know this has occurred. This appears to be the case for many large language models (LLMs).[246] Through his own usage of the systems, Volokh has noted that they 'seem to routinely erroneously produce false and defamatory statements'.[247] For some reason, these hallucinations have the proclivity to be reputationally damaging and of a particularly destructive nature to personal dignity. On testing, ChatGPT for example regularly accuses people of crimes they have not committed.[248] There is currently a libel suit in progress against OpenAI as well as a claim open against search engine Bing, which uses AI program ChatGPT 4, which falsely claimed that a professor was a convicted criminal (after the program confused the professor for a known

[243] Ibid 603–4.
[244] Cheong, Caliskan and Kohno (n 237) 6.
[245] Henderson, Hashimoto and Lemley (n 240) 591.
[246] Volokh (n 234) 492.
[247] Ibid.
[248] Henderson, Hashimoto and Lemley (n 240) 591.

terrorist with a similar name).[249] Quotations purportedly from others communicated to users by AI models are regularly incorrect or completely fabricated.[250] Such hallucinations do not appear to be attributable to 'bad inputs': it is rather 'the way large language models work'.[251] Early research suggests this may be because these models generate predicted text, which leads to the AI system falsifying information in order to provide an answer to a question prompt in the event the program is not able to (honestly) answer the question.[252] False statements such as these are presented conclusively as accurate information by these language models.[253] Cheong, Caliskan and Kohno have even noted that AI can be 'opinionated' – thereby unknowingly influencing a user to believe that every answer generated is implicitly correct.[254] As many individuals have moved to consuming their news through social media rather than traditional newspapers, communication channels have now changed and AI is looking poised to dominate online social media in the near future.[255] The hallucination problem is so rife that AI language models have been seen to generate defamatory outputs even to non-malicious questions about an individual.[256] This defamation problem is only likely to increase in significance as people across the globe begin to rely on AI tools and implicitly believe the information generated by these systems.[257]

These practical threats are underscored by the current lack of legislative or adequate common law control over AI

[249] Volokh (n 234) 492.
[250] Ibid 529.
[251] Henderson, Hashimoto and Lemley (n 240) 592.
[252] Ibid.
[253] Ibid.
[254] Cheong, Caliskan and Kohno (n 237) 3.
[255] Albright (n 174) 122–3.
[256] Henderson, Hashimoto and Lemley (n 240) 596.
[257] Volokh (n 234) 493.

developments on a significant scale. The most comprehensive piece of legislation to date concerning AI worldwide is the European Union's recent and pioneering 'AI Act', although this is not UK law after 'Brexit' in 2020. Despite the scale of AI's insurgence into everyday lives online, the UK appears reticent to apply strict regulation to this online space, regardless of the obvious threats of AI programs to personality interests and human rights.[258] It is not clear if this legal reticence is due to a lack of understanding, fears about quashing economic or creative benefits, or instead a concern over the complexity of the legal task involved. Compounding the problem, there seems to be little appetite on the part of AI program manufacturers to prioritize fixing the problem of AI-generated defamation, despite its prevalence.[259]

c. Who should be responsible for an AI tool's defamatory speech?

It is argued here that there is a clear and pressing need for the law to accommodate claims in defamation when an individual is defamed by an AI program and then these defamatory statements are published – in other words, communicated[260] – to another individual. This is for two reasons, outlined earlier. One is because this is a very real threat, as demonstrated by the propensity of LLMs to defame, shown through their proclivity to hallucinate. The other is that defamatory imputations spread in this way can cause very real harm for the individuals concerned, as information is now regularly consumed through the medium of the internet, with a degree of trust extended to AI programs that are marketed as reputable. Despite legislative

[258] Cheong, Caliskan and Kohno (n 237) 2.

[259] Volokh (n 234) 493.

[260] English law has made clear that publication, even to only one person (outside the claimant themselves), will be sufficient publication, or as Horsey and Rackley term it, 'communication', in the law. See Kirsty Horsey and Erika Rackley, *Tort Law* (7th edn, OUP 2021) 504.

reticence, calls for legal involvement in this area are becoming louder. Albright has argued that a 'stricter duty' in US law should apply to those who use an AI device to defame others, in order to tackle burgeoning risks.[261] Due to the increased possibilities the internet provides to defame, aided by AI programs, the rate at which defamatory information can be generated and communicated online is unfathomable. It is crucial, then, that there are routes for redress in defamation against both third-party publishers of that information but also AI program manufacturers. For example, a third-party publisher may be a living individual who uses an AI program to generate defamatory content; the program then publishes it further afield, in a *defamation by AI tool* scenario. A manufacturer would be the company responsible for designing, programming, training and putting the AI program that has generated that defamatory content to market. Of course, an AI program itself 'publishes' or sufficiently communicates the defamatory statement when it answers a text prompt from a human user and defames another individual. The requirement of publication in English defamation law has always been satisfied by one-to-one communication, outside of (for example) a call between a claimant and defendant themselves.[262] It is not currently legally practicable to bring a claim against an AI tool itself in the English jurisdiction – firstly, because such a tool does not have a separate legal personality in the eyes of English law, unlike companies;[263] and secondly, because it would not

[261] Albright (n 174) 105.

[262] Henderson, Hashimoto and Lemley (n 240) 635–6; Volokh (n 234) 504; and Horsey and Rackley (n 260) 504.

[263] Although the advantages of effectively 'incorporating' AI and therefore giving it legal personality are being mooted. See, for example, James Russell, 'Artificial intelligence and separate legal personality' (*Inside Tech Law*: *Norton Rose Fulbright*, 12 November 2019) www.insidetechlaw.com/blog/2019/11/artificial-intelligence-and-separate-legal-personality accessed 18 December 2024; and Visa AJ Kurki, 'The legal personhood of artificial intelligences' in *A Theory of Legal Personhood* (OUP 2019) ch 6.

independently have the funds to support any compensation awarded in a successful claim. Website hosts that merely host defamatory content generated by a third party (including an AI program) may be protected under the s 5 operators of websites defence in the Defamation Act 2013, discussed earlier in this chapter. However, this defence would likely fall away if the website itself played an active rather than passive role in disseminating defamatory information generated by AI software.[264] The defence would also fail to be operational in the event that the web hosts had been made aware of the defamatory content generated by an AI tool hosted on their website, were effectively 'on notice',[265] and yet failed to act on that information as per the statutory instrument.[266] Section 5 will not provide a defence for the manufacturer or user of an AI tool, as the type of AI program known to commonly defame is generative and therefore is creating the defamatory information itself – and not merely regurgitating or 'hosting' information from elsewhere on the internet. Similarly, the US defence in s 230(c)(1) of the Communications Decency Act that protects those behind computer services from liability in defamation (in the event that the information had been provided by another content provider), which includes social media and search engines, is not thought to extend to generative AI systems.[267]

It is argued here that companies producing generative AI programs should be subject to liability in defamation for defamatory outputs of such tools. In addition to the points mentioned, there are two further reasons for this liability regime. Firstly, there may be a lack of recourse for claimants

[264] As s 5 is a defence for a web host who has *not* posted the statement complained of, as stipulated in s 5(2) Defamation Act 2013.

[265] For an argument to this effect in the US context, see Henderson, Hashimoto and Lemley (n 240) 647.

[266] The Defamation (Operators of Websites) Regulations (n 164).

[267] Henderson, Hashimoto and Lemley (n 240) 620–2 and Volokh (n 234) 494.

in terms of who else to sue – in the event that a third party has communicated information generated using an AI tool, but is not contactable. Secondly, an AI program itself is not currently able to be claimed against in English law, due to the abovementioned reasons of lack of separate legal personality and funds. Holding AI software *manufacturers* as accountable instead offers security for claimants, as such companies may have 'deep pockets' and are therefore well positioned to compensate claimants for the affront to personal dignity caused. Holding manufacturers as liable is also beneficial as they will likely have the power and influence to amend AI programs in the future, to discourage defamatory statements being generated. Indeed, Henderson, Hashimoto and Lemley have argued that holding manufacturers accountable provides a valuable incentive for them to improve the offering that they provide in a way which is beneficial for society at large.[268] If more informational safeguards are incorporated into an AI program, the less likely it is to generate harmful defamatory content in the future. This can be done through supporting programs with more rigorously curated input data, the inclusion of more checks and balances into the software and an increased amount of human testing to minimize hallucinations.[269] Cheong, Caliskan and Kohno have observed that there are insufficient 'market incentives' at present to persuade developers to produce higher-quality AI programs designed with limiting defamation in mind. This is because 'profit incentives do not automatically encourage robust safety efforts'.[270] English courts finding AI program manufacturers accountable in the *defamation by AI tool* scenario would go some way to encourage best practices in the industry. Finally, the famed English tort law policy argument of 'floodgates', or the courts wishing to safeguard

[268] Henderson, Hashimoto and Lemley (n 240) 636–7.
[269] Ibid 648–9.
[270] Cheong, Caliskan and Kohno (n 237) 6.

against uncontrolled liability, is not sufficiently engaged here as to cause the judiciary to feel there is any pressing concern in extending the law in this way. As has been argued in the previous chapter, English defamation law is already difficult to successfully argue as a claimant in the digital age – it has many hurdles, such as the serious harm threshold, reference to the claimant and publication itself, and a plethora of defences that a defendant could potentially rely on. Claims against defendant manufacturer companies would be limited by all of the abovementioned factors and therefore development of the law in this way would not result in an insurmountable deluge of claims.[271] This is not to say that it is beyond the capabilities of English common law to construct a method for the AI system itself to be individually responsible in defamation, in the same way that an incorporated company can be responsible in various contexts, in the realms of commercial law – rather, it is a question of practicalities, the status quo in the law and the ability of claimants to obtain a meaningful remedy.[272]

If companies producing generative AI programs are to be found capable of liability in English defamation law, the question then arises of what companies can do in order to avoid liability. One potential recourse to bolster the reliability of AI tools and ensure they do not defame is to make sure the software rigorously fact-checks itself or its sources. An appropriate standard of reliability would be whether 'the algorithm's methods meet the standards of journalistic procedure'.[273] The often limp disclaimers employed by AI companies in an attempt to shield themselves from legal responsibility should not be sufficient to stop an action against them in defamation – as a short disclaimer stating that information may not be entirely accurate is not the same as

[271] For this argument as broadly stated, ibid 12.
[272] For an argument relating to practicalities, see Volokh (n 234) 508.
[273] Albright (n 174) 117.

clearly stating that something is fiction or parody (ensuring that nothing is believed by an audience), particularly as AI devices are partly marketed on their reliability.[274] A practical route to make sure generative AI tools provide users with accurate information is for developers to dedicate time for the system to learn rigorously from 'human feedback' and other training modifications. This allows human developers to curate data to follow certain overarching principles.[275] This process is already engaged in to some extent by AI manufacturers. However, even if this approach is used in the development process, Henderson, Hashimoto and Lemley observe that achieving only 65 per cent accuracy is still a normal percentage to be expected,[276] which is concerningly low. To compound this problem, austere content moderation like this can sometimes *take away* from accuracy, in a confusing quirk of development.[277] There are other simpler solutions posited – such as ensuring that an AI program refuses to answer a user prompt if it cannot generate an accurate answer – which can reduce the need to AI to hallucinate in order to find one.[278] In essence, to shield itself from liability for defamation published by its AI tool, a manufacturing company should have to demonstrate that 'it has taken all reasonable measures to prevent the propagation of harmful statements',[279] in a standard reminiscent of common law negligence in the English system. Volokh has noted that this is similar to holding a newspaper editor as liable if they have not robustly fact-checked an article that then defames someone,[280] and in such a circumstance the s 4 publication on a matter of public interest defence in English law would

[274] Volokh (n 234) 500–2.
[275] Henderson, Hashimoto and Lemley (n 240) 612.
[276] Ibid 613.
[277] Ibid 615.
[278] Ibid 619.
[279] Cheong, Caliskan and Kohno (n 237) 10.
[280] Volokh (n 234) 522–3.

likely fall away.[281] Particular attention needs to be paid to AI devices attributing false quotations to individual, when in fact the quotations have been hallucinated by the device.[282]

It is clear from this discussion that AI tool manufacturers have a mountain to climb to ensure accuracy of AI tools. This is a costly and time-consuming process, the parameters and specifics of which are not yet fully understood. In the meantime, if AI manufacturing companies are prepared to accept the large amounts of profit generated by popular AI tools,[283] they must also accept the liability that should ultimately arise through such devices defaming third parties.

d. Deepfakes and defamation

The arguments rehearsed apply to generative AI in a comprehensive manner. Before this chapter closes, it is important to specifically address a particularly harmful branch of AI development: the rise of 'deepfakes' on the internet. A deepfake can take the form of (for example) a video that appears to be of a real person who is speaking, such as a politician – however, although the video appears to be real it is fact fictional and generated by AI. A deepfake could also take the form of a modified photograph of a person in a compromising position – perhaps sexual – or otherwise implying reputationally damaging behaviour or beliefs. This is another way that an individual can be *defamed by an AI tool*. Deepfakes are particularly damaging to a person's psychological

[281] As the grounds of 'reasonable belief' that publishing something was in the public interest would not be met, in s 4(1)(b) of the Defamation Act 2013.

[282] Volokh (n 234) 522–6.

[283] OpenAI was valued at $157 billion US dollars in 2024. See David Curry, 'ChatGPT revenue and usage statistics (2024)' (*Business of Apps*, 13 November 2024) www.businessofapps.com/data/chatgpt-statistics/ accessed 20 December 2024.

integrity, personal dignity and self-perception, as being exposed to a lifelike video of something they appear to be doing may even lead them to question themselves. High-quality deepfakes, many of which can be seen today due to advancements in technology, appear to be real and can be incredibly convincing. The defamatory implications of such videos are obvious and deepfake production has been something of an internet phenomenon. Deep learning is responsible for this breakthrough,[284] and the production value of deepfakes is increasing as individuals share best practises for producing deepfakes through internet message boards.[285] The AI tools that power the creation of such fakes are often open source and there are applications available to download to create deepfakes that 'require little to no coding skills'.[286] This means that the creation of deepfakes is possible by many and not limited to a class of individuals who necessarily have the technical skill or financial backing. A shocking 90 per cent of deepfakes are pornographic.[287] This means that deepfakes pose a particularly concerning threat to one's personal dignity, as when one's likeness is misused in this intimate and violating manner it is also likely to cause feelings of shame and embarrassment and have a potentially grave impact on individual reputation. The disturbing motive of creating deepfakes of real individuals depicting sex acts online is clear, as Karasavva and Noorbhai observe: 'using deepfake technology, anyone could direct their own pornographic material, casting people from their own lives'.[288] To make matters worse, accurate detection of

[284] Ángel Fernández Gambín, Anis Yazidi, Athanasios Vasilakos, Hårek Haugerud and Youcef Djenouri, 'Deepfakes: Current and future trends' (2024) 57(64) *Artificial Intelligence Review* 1, 1.

[285] Vasileia Karasavva and Aalia Noorbhai, 'The real threat of deepfake pornography: A review of Canadian policy' (2021) 24(3) *Cyberpsychology, Behavior, and Social Networking* 203, 205.

[286] Ibid.

[287] Ibid 203.

[288] Ibid 204.

whether material is a deepfake is complex and difficult. It is particularly challenging as often videos are partially real, but partially tampered with, and as such are 'augmented reality'. Current deepfake detectors find it challenging to spot this nuance – as detectors usually classify videos in binary fashion as entirely real or entirely fake.[289] Detectors also find identifying deepfakes a challenge as deepfake manipulation can be of varying types, such as auditory, visual or both.[290] Studies have been undertaken which show that (unhelpfully) social media platforms can make it *more difficult* to detect deepfakes – so if one is *defamed by an AI tool* in the form of a deepfake and then it spreads online through social media (leading to a crossover with the *defamation by social media* scenario) it can be even more challenging to find redress. Gambin et al. explain: 'some manipulations are performed by social media networks before uploading any content. This is known as social media laundering and removes clues with respect to underlying forgeries, and eventually increases false positive detection rates.'[291]

Blockchain technology shows promise in terms of developing more accurate deepfake detection mechanisms in future.[292] Requiring 'proof of authenticity', or proof of source, can establish which videos are real or deepfakes, and blockchain allows the transfer of information to be highly controlled and verified at each stage, through 'transactional transparency'.[293] So-called smart contracts can also be used to determine who handles a video at any given time, to help ascertain if a video is real – and if not, who has altered the video to create a deepfake, so that they may individually be actioned against in defamation law or otherwise. To more comprehensively

[289] Gambín et al. (n 284) 12.
[290] Ibid.
[291] Ibid 13.
[292] Ibid.
[293] Ibid 16.

tackle the spread of deepfakes online, social media sites must be proactive and champion testing new ways to combat the rapid spread of disinformation on their platforms.[294] If social media sites robustly safeguard against the spread of deepfakes this may reduce the need for defamation law as an imperfect medium through which claimants may seek recourse. A broad mobilized response to deepfake production is necessary in order to tackle their wide spread – corporations, the media and governments must work together to recognize their threat to personality interests, human rights and human dignity.[295]

A further question is what area of law is best suited to tackle the emergence of deepfakes. As argued, defamation using deepfake technology is yet another form of *defamation by an AI tool* and English defamation law is a possible recourse through which claimants may attempt to seek redress. However, this is far from an ideal solution, as the requirements to make out an action in defamation (a defamatory statement, the serious harm threshold, reference to the claimant, publication) are many and varied and subject to a broad array of defences. There is also as yet little case law in this area, so it is unclear how sympathetic the common law will in fact be to such a claim, despite the powerful arguments in favour of liability rehearsed earlier in this chapter. Given that the harm caused by pornographic deepfakes is so severe, it is more appropriate that criminal law regulates this area. The UK's recent Online Safety Act 2023 criminalizes the sharing of 'intimate' deepfake images.[296] The police encourage people to report intimate deepfakes, so that those sharing these

[294] Ibid.

[295] Ibid.

[296] Section 188 Online Safety Act 2023, amending s 66B Sexual Offences Act 2003. For further reading see 'Criminalising deepfakes: The UK's new offences following the Online Safety Act' (*Herbert Smith* Freehills, 21 May 2024) www.herbertsmithfreehills.com/notes/tmt/2024-05/criminalising-deepfakes-the-uks-new-offences-following-the-online-safety-act accessed 20 December 2024.

images can face prosecution.[297] Further criminal legislation on this issue is expected, with some additional forthcoming legislation 'washed up' in the wake of the November 2024 UK general election.[298] The recent Data (Use and Access) Act 2025 includes further provisions to combat deepfakes in clause 138, providing that it is an offence to create,[299] or intentionally request the creation of, a 'purported intimate image of another person (B)' without B's consent. This amends the Sexual Offences Act 2003.[300] It is argued here that regardless of the fact that *defamation by an AI tool* should be actionable under English defamation law, it is important that the very serious harm rendered by pornographic deepfakes is dealt with through the medium of criminal prosecution, befitting the severity it deserves as a sex offence. There seems to be a strong appetite among the public in many countries to deal with this issue using criminal law means.[301] That is not to say that defamation law does not have a place in regulating deepfakes; it may be more appropriate for an action regarding a deepfake that is reputationally damaging but not pornographic. For the reasons outlined in section 'c. Who should be responsible for an AI tool's defamatory speech?', AI program manufacturers who produce

[297] 'Deepfakes: Reporting it to us' (*Police.uk*) www.police.uk/advice/advice-and-information/deepfakes/deepfakes/deepfakes-report-it/#:~:text=It's%20illegal%20to%20share%20or,report%20this%20to%20the%20police accessed 20 December 2024.

[298] That went towards criminalizing the creation of such images. See Ministry of Justice and Laura Farris, 'Government cracks down on "deepfakes" creation' (*GOV.UK*, 16 April 2024) www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation accessed 20 December 2024.

[299] See Data (Use and Access) Act 2025 clause 138. This amends the Sexual Offences Act 2003 accordingly. See https://www.legislation.gov.uk/ukpga/2025/18/enacted accessed 23 July 2025.

[300] Ibid

[301] Matthew B. Kugler and Carly Pace, 'Deepfake privacy: Attitudes and regulation' (2021) 116 *Northwestern University Law Review*, 611, 611.

tools with which to create deepfakes should be potentially liable in defamation law, as well as publishers of the content. In relation to this type of technology, the argument that a manufacturing company had 'done enough' to ensure accuracy of output data of an AI tool falls away when a program is specifically designed to create deepfakes – as defamation is overtly likely, rather than an unintended consequence. Many tools that optimize the creation of deepfakes are already in existence. As Karasavva and Noorbhai observe: 'Scraper or DownAlbum allow users to download all pictures and videos uploaded on publicly available Instagram and Facebook accounts. Thus, using these tools one can easily create the datasets necessary to train the deepfake algorithm'.[302]

It may also be possible to use the aged tort of *Wilkinson v Downton*,[303] which affords liability for the intentional infliction of emotional harm as an alternative route to redress in tortious deepfake cases. However, more recent decisions on the *Wilkinson* rule, such as *Wainwright* and *Rhodes*,[304] have set claimants a high bar of a conduct, mental and consequence element to be overcome.[305] In any event, it is outside the scope of this work to discuss this further.

III. Concluding remarks for Part II

This part of this chapter has considered two technological eventualities of online defamation: third-party comments containing defamatory material in the *third-party poster* scenario and *defamation by AI tool*, be it by a generative LLM hallucinating or the more malevolent method of a reputationally damaging deepfake. Both of these are new scenarios that have been

[302] Karasavva and Noorbhai (n 285) 204.
[303] Ibid 206 and *Wilkinson v Downton* [1897] EWHC 1 (QB).
[304] *Wainwright v Home Office* [2003] UKHL 53 [46]–[47].
[305] *Rhodes v OPO and Another* [2015] UKSC 32 [73]–[87].

brought to the fore by technological advancements of the digital revolution since the early 2000s, and both scenarios pose unique legal challenges for the traditional area of English defamation law. The prevalence of defamation by AI and the sheer number of third-party comments on social media (and other) websites has increased the defamatory potential of the internet exponentially. It has been argued that the s 5 defence in the 2013 Act provides a powerful incentive for web hosts to act in accordance with the relevant statutory instrument, providing a fair balance between the interests of those defamed online seeking redress and web hosts themselves, as well as free expression online. The approach of the 2013 Act in this regard must be commended, although there are still outstanding issues posed by s 5 (such as whether social media websites are caught in its remit and the fact that no interim relief pending trial is available for potentially defamatory third-party posts). Further, it has been argued here that there is no clear reason why reputational damage through an *AI tool* should not be taken seriously under English defamation law. In order to ensure redress, AI tool manufacturers must be held legally responsible in English defamation law for creating an AI program that defames, subject to requisite demonstration of rigorous accuracy checking on the part of manufacturers during product development. Finally, deepfakes are a deeply concerning phenomena that pose some of the most potentially grievous potential to reputation, personal dignity, self-perception and human rights. While defamation law is an avenue that could be pursued to combat these creations ex post, thankfully English criminal law has begun to criminalize sharing the most harmful types of deepfakes, those of a sexual nature.

★★★

Conclusion for Chapter 3

Chapter 3 as a whole has argued that a number of different changes have made it uniquely difficult for claimants who are

the victims of online defamation to bring a successful action in England and Wales. Firstly, the law was partially reformed by the Defamation Act 2013, just over a decade ago at the worst possible time. Online defamation was beginning to gain pace, yet the 2013 Act's sympathies primarily lie with defendants and the Act openly prioritizes freedom of expression at the expense of reputation. Part I focused on the introduction of ss 1 and 8 Defamation Act 2013 as two examples of changes the law brought in that particularly disadvantage claimants in online defamation cases. Secondly, in Part II, this chapter considered the defence for 'host' websites in s 5 of the Defamation Act 2013 as a (lone) example of good practice introduced by the Defamation Act 2013 and placed it in the context of Strasbourg jurisprudence. Finally, it was argued here that defamation law can and should do more to protect claimants defamed by AI tool, a scenario rendered increasingly likely due to the burgeoning widespread use of generative LLMs.