CarnegieUK TRUST

CHANGING MINDS • CHANGING LIVES

The Carnegie Statutory Duty of Care and Fundamental Freedoms

Professor Lorna Woods (University of Essex)

December 2019

1. Introduction

Carnegie UK Trust proposed that a statutory duty of care be imposed on social media platforms so that the online environment might be improved for users.[1] The Carnegie proposal focusses on the underpinning social media service in terms of its design and business operations. This differs from an approach which seeks to specify particular items of content as problematic and sees solutions only through the lens of take-down. This latter approach is the traditional publisher model which operates on the basis of a distinction between publisher, who has a role in the creation/selection of content, and the intermediary, who knows nothing of the content. Such a binary approach, we have argued, is inappropriate for the internet context, specifically as regards social media. The Carnegie proposal aims to tackle the abuse-enabling environment that some platforms seem to have become by focussing on the question of whether service providers have adequately taken into account the risks of the way their systems have been designed and run – whether this be through design features influencing content or nudging users to certain sorts of behaviour at the point of creation or engagement (for example through reward mechanisms); or in the context of how content is discovered (for example through the use of algorithms to promote and recommend content). Takedown would be a mechanism of last resort – though in some instances may well be appropriate (consider, for example, the need to tackle child sexual abuse and exploitation material).

Inevitably, there is a connection between the duty and online content. A natural concern is how such an obligation might impact fundamental freedoms, including but not limited to freedom of expression. The purpose of this piece is to consider that intersection and specifically whether concerns about human rights preclude the imposition of a duty of care as envisaged in the Carnegie paper. It first outlines the human rights framework from a British perspective taking into account the impact of the European Convention on Human Rights (ECHR) and the case law on that. It then considers how the various design techniques and business choices could be assessed from a rights-based perspective.

2. Overview of Human Rights

A number of rights could be implicated by the Carnegie statutory duty of care, notably freedom of expression (Article 10 ECHR), but also the right to religion (Article 9 ECHR); the right to private life (Article 8 ECHR); freedom of association (Article 11 ECHR); and the right to education in Protocol 1, Article 2. Article 14 prohibits discrimination in the enjoyment of rights. By contrast to the position in the US, the European Article 10 right to freedom of expression (and other rights including religion, association and private

life) are limited. This does not, however, give States carte blanche to interfere with rights. Instead, any government action must satisfy a three-stage test: that the interference seeks to achieve a legitimate aim (as listed in the Convention); that the interference is in accordance with the law; and that it is necessary in a democratic society.

Where the State sets up a regulatory framework within which the private actors must operate, it seems likely that the framework itself and the actions of the regulator, if not the decisions of the platforms themselves, trigger the consideration of Convention rights. It should be acknowledged that the choices of platforms, as private actors, do not automatically engage Convention rights; rights are addressed to the signatory states. While this may seem like a neat boundary, in practice matters are not that clear. It does not mean that a State never has responsibility for the actions of private actors, nor should it be seen as suggesting that private actors, particularly transnational corporations, need have no regard to human rights. The responsibilities of companies are elaborated at the UN level in the non-binding[2] Ruggie Principles.[3] Because the objective of this paper is to look at whether human rights considerations constitute a barrier to the duty of care model, the Ruggie principles are not dealt with directly in this analysis. Nonetheless some of the arguments made here, may also be relevant to that context. There are three levels of rights considerations:

1. where the State acts, including the actions of a regulator set up by the State even if that regulator is independent;
2. actions carried out by private actors but required by the State; and,
3. actions carried out by private actors but not required by the State.

In practice, it may be hard to differentiate between categories (2) and (3) in a context in which a State requires some action.

In relation to the first set of considerations, as a matter of domestic law all public bodies, and this would include a regulator set up under such a framework, are bound by the Convention rights;[4] public bodies also have obligations under the Equality Act 2010.[5] Any regulatory framework must allow for the appropriate balancing of such rights, and in this freedom of expression – as recognised in the case law of the Supreme Court in *Re S (A Child)*[6] – does not take automatic priority, and in *PJS*[7] the Supreme Court clarified that s.12 Human Rights Act, which limits the granting of *ex parte* relief, does not give Article 10 rights priority over Article 8.[8]

In relation to the second level, it would seem that where private actors are required by law to intrude into the rights of others, that action could be attributed to the State. Within the duty of care model, there is a significant degree of freedom of choice in terms of individual design and business choices, so it is less clear whether such an attribution could be made. Much may depend on the advice given through recommended best practice and the requirements, for example through codes, set down by the regulator.

Another difficult question for a rights-based analysis is the distinction between the freedom from interference and the right to do something or to make claims on others. In some instances, a State may be under an obligation to intervene in disputes between private parties to ensure that rights are not illusory. States may have an obligation to protect individuals in some contexts – and it may be that the list of harms can been seen from this perspective. It would be hard to argue, however, that a user had a right to use a particular social media platform based on freedom of expression (see the case of *Appleby*[9])

– though this could change were there to be no alternative platform available. In any event, platforms should not discriminate by reference to protected characteristics. In principle, as recognised by the Council of Europe recommendation, this means that platforms are free to limit speech (and other rights) on their platforms in accordance with their respective terms of service if they so choose, especially given that the platform's own rights might come into play (see further below) – though they may choose to take into account the Ruggie principles or regional equivalent documents. This freedom to set their own standards means, for example, some sub-reddits have stringent moderation;[10] in Mastodon, different communities adopt very different stances as to what is acceptable on their respective instances (contrast for example the instance named Toot Cat with its Code of Conduct[11] with the UK instance.[12])

This question may become yet more complex should we seek to argue that a right to freedom of expression includes the rights to access particular communications tools (e.g. augmented reality), especially where the platform would not otherwise choose to deploy those tools. Even for States, there is a distinction between an interference with a right and an obligation to maximise the possibilities for the exercise of the right. This must be the more so in the case of obligations on private actors.

For the sake of argument, in the following discussion we will generally assume that rights-based claims could be made – though it should be noted that this may not be the case in all circumstances.

2.1 Freedom of Expression

The protection awarded to expression is broad, and the Court has taken a very wide approach to the question of what speech is and who may exercise the right. By contrast to the UN ICCPR, the ECHR allows companies to claim rights. Typically, a high level of protection is accorded to political speech[13]and the activities of the media.[14] While the right to freedom of expression is important – and famously includes information and ideas that offend, shock or disturb[15] – as well as allowing the speaker to choose the form of speech[16], it is not unlimited. Indeed, some speech may fall outside the protection of the Convention altogether by virtue of Article 17 ECHR, which prevents persons from relying on Convention rights to undermine the values on which the Convention is based[17]. Factors to be taken into account when considering whether an interference with the right to freedom of expression under Article 10(2) include: the content of the speech – with the limits of permissible criticism being wider for government and public actors than for private citizens;[18] the speaker (with the media being highly protected); the nature/size of the audience; the means of transmission; the nature of the penalty/ intrusion; and the aims the State seeks to achieve (as listed in Article 10(2) and in the last sentence of Article 10(1) ECHR).

Freedom of expression applies not only to the truth, but value judgments and opinion – though even for the media there must be some form of connection between existing 'facts' and the opinion. In *Medžlis Islamske Zajednice Brčko*, the Grand Chamber summarised its jurisprudence in this area noting that "special grounds are required before the media can be dispensed from their ordinary obligation to verify factual statements that are defamatory of private individuals". This assessment depends on the facts, including "other elements such as whether the newspaper had conducted a reasonable amount of research before publication" and "whether the newspaper presented the story in a reasonably balanced manner" (para 108).[19] In *E.S* (discussed below) a contributory factor was that the speaker relied on "untrue facts" (para 53).[20] While politicians enjoy latitude because of the importance of political speech, they are not automatically exempt from the rules relating to hate speech even when speaking in Parliament (see e.g. *Pastors*[21]).

Artistic speech is recognised as a category worthy of protection, including comedy and satire,[22] where the Court has introduced the test of what a reasonable reader would understand to be satirical (taking into account the context of the speech).[23] In *Sousa Goucha,*[24] the Court emphasised that in its jurisprudence on the topic, it is significant that 'the artistic creations in question were made against a background of political critique and debate'. Nonetheless, the reliance on humour is not a 'get out of gaol free' card[25], especially when the speech constitutes hate speech, as in the case of *M'Bala M'Bala.*[26] Indeed, in the context of artistic speech, States have been found to be able to restrict speech in the interests of morality and religious sensibilities.

In the controversial case of *ES,*[27] which concerned the fine of a woman who described the Prophet Muhammad's marriage to Aisha as paedophilia, the Court noted that freedom of expression carries responsibilities. In the Court's view "[a]mongst them, in the context of religious beliefs, is the general requirement to ensure the peaceful enjoyment of the rights guaranteed under Article 9 to the holders of such beliefs including a duty to avoid as far as possible an expression that is, in regard to objects of veneration, gratuitously offensive to others and profane".[28] Freedom of religion is, of course, another fundamental right, albeit not an unlimited one, and States have a wider margin of appreciation when balancing conflicting rights. In *ES*, the Court noted that States "have the positive obligation under Article 9 of the Convention of ensuring the peaceful co-existence of all religions and those not belonging to a religious group by ensuring mutual tolerance".[29] This links back to the need for tolerance and for pluralism that the Court has suggested is necessary to support democracy. Of course, freedom of religion should not automatically override speech rights; nor can it be used to justify hate speech (e.g. *Belkacem*[30]).[31]

Freedom of speech may come into conflict with other rights – notably the right to private life (Article 8). There is long standing jurisprudence from the Court on how to balance freedom of speech (particularly when exercised by the media) with the right to privacy with the Court identifying (see e.g. *Axel Springer*[32]) a range of factors to take into account, including whether there is a contribution to a topic of public debate; the notoriety of the person seeking to exercise Article 8 rights; the prior conduct of that person in relation to the media as well as the content, form and impact of the story.[33]

2.2 Right to Private Life

The right to private life is sometimes described as being the right to privacy. While it certainly includes the right to privacy (including data privacy and freedom of state surveillance), Article 8 ECHR is far broader than that. It includes the right to moral, physical and psychological integrity. Issues relating to abortion, homosexual and transgender rights, medically-assisted procreation, and even personal appearance, for example, all fall within its scope. The Court has held that the "preservation of mental stability is in that context an indispensable precondition to effective enjoyment of the right to respect for private life".[34]
In an online environment, Article 8 may be engaged where platforms are keeping records of a user's pattern of use, as well as by the sharing of private information or by the harassment of an individual, especially a vulnerable individual. Further, Article 8 includes individuals' identity as part of a group (e.g. the dignity of the victims and the dignity and identity of modern-day Armenians in the context of the mass deportations and massacres suffered by the Armenians in the Ottoman Empire in 1915 and beyond in *Perinçek*[35]). Conversely, the Court has also accepted that the limit of permissible criticism may be broad where those commenting are well-known professional journalists and the matter in issue one of public interest. When the right to reputation is at stake, even for private persons, the attack must attain a certain level of seriousness.[36]

As with Article 10, States may be under positive obligations as regards Article 8. This includes a duty to maintain and apply in practice an adequate legal framework affording protection against acts of violence by private individuals (see e.g. *Soderman*[37] concerning the attempted covert filming of a 14-year-old girl by her stepfather while she was naked). As regards cases of rape and sexual abuse of children (e.g. *M.C v Bulgaria*[38]), which implicate 'essential aspects of private life', the State's positive obligation includes enacting efficient criminal-law provisions and ensuring the effectiveness of a criminal investigation (e.g. *KU v Finland*[39]). A case has recently been communicated which involves the failure of the State to protect the complainant from repeated acts of online violence and the republication of nude photographs as well as the creation of fake social media accounts.[40] In principle, the failure to take action in relation to school bullying falls within Article 8, but there must be some basis on which State authorities should be aware of the risk – vague and general complaints do not suffice.[41] This could be seen to suggest that a positive obligation arises in relation to specific individual(s), rather than an obligation in relation to an at-risk group in general. Nonetheless, the Court has held that there should be a legal framework criminalising anti-minority demonstrations and which should afford effective protection against harassment, threats and verbal abuse.[42] For less serious matters, civil law remedies may suffice – though a would-be claimant is not guaranteed to win his or her claim in all circumstances (see e.g. *Tamiz v UK*[43]).

2.3 Discrimination

The right to non-discrimination under the Convention is not a free-standing right but must link to one of the other rights guaranteed by the Convention (though it is not necessary to show a violation of that right). Where there is a violation of the substantive right, quite often the Strasbourg Court contents itself with this finding and rules that there is no separate issue as regards Article 14. When a mob carried out a pre-planned attack on the homes of some Roma individuals based on anti-Roma sentiment which resulted in the Romas' displacement, the failure of the police to prevent the attack was found to constitute a violation of Article 14 in conjunction with Article 8.[44] In *Opuz,*[45] the Court ruled, in a claim arising from domestic violence abuse and murder, that Article 14 fell to be interpreted in light of the specialist jurisprudence of the Convention on the Elimination of all Forms of Discrimination against Women, which jurisprudence recognised violence against women, including domestic violence as a form of discrimination against women. In the report of the UN Special Rapporteur on Violence Against Women on online violence against women and girls, it was noted that online and internet-facilitated forms of violence against women have become increasingly common, particularly with the use of social media platforms and other technical applications, and are used to maintain and reinforce patriarchal norms.[46] The impact of discrimination in the context of silencing others, although increasingly recognised,[47] has not been developed by the court (note deference to value ascribed to freedom of expression in *Sousa Goucha*).

The online environment may also lead – through 'personalisation' and targeting of content – to discrimination in terms of the information that people receive, which may also be discrimination in the context of Article 10.[48]

There is one final aspect to discrimination: that is the rights of those with disabilities and their ability to access services.

2.4 Approach to the Internet

The case law of the court with regard to the internet is not yet well-developed. Most of the cases have to date mainly focussed on content (whether in the context of criminal penalties imposed on speakers[49]or civil actions for content[50]). The Court's approach is based on its existing jurisprudence understood in modern conditions, and gaps in protection should not arise because of the technology used. This equalising impetus means that the publication of false information would not be protected just because it is posted on the internet. While *Tamiz* accepted that 'vulgar abuse' was common on internet forums, the Court has on a number of occasions accepted that in principle the positive obligation to protect Article 8 applies also in the context of third party comments on blogs.[51] Nonetheless, the special position of the media seems to hold good on the internet too: the Court has held that acceptable criticism is broadly drawn where those commenting are professional journalists well-known to their audience and who are commenting on matters of public interest.[52]

Another issue is that of collateral censorship which arises if a blocking order is too broadly defined.[53] So, a takedown notice for a specific item of content has fewer ancillary effects than, for example, an order blocking YouTube which has both problematic and harmless content on it – though to complain, users must be able to show that they were specifically affected so as to be able to prove that they were victims for the purposes of the right of individual petition under the Convention. While a prohibition on the broadcasting of a particular item of content would similarly affect large numbers of people (not just the speaker but the potential audience), the issue is probably more problematic in the context of the internet because of the size of some of the platforms.

The Court has noted specific features of the internet. The internet is particularly important for the media in its duty to impart information and ideas and to create archives via the internet and also has an important role to play in facilitating access to information. Risks to other rights (notably Article 8), however, also increase because of the viral nature of the internet (certainly as regards some types of content) and its amplifying effect – it is possible that information may be used in a way not foreseen. Because of this, journalistic ethics become more important in the light of an audience's right to information. This is particularly important where minors are concerned. There are some cases which address situations specific to the internet – notably the role of those hosting content from others. In *Delfi*, the Court found that the fact that national courts held the finding of Delfi liable for defamatory comments posted by a user did not constitute a violation of Article 10 – in part because the website (a professional site) had to some extent encouraged the content. In subsequent cases, violations of Article 10 were found because the language used was less extreme (e.g *MTE*). The Court has taken a similar approach in relation to claims under Article 8 or Article 10, and in addition to the factors used for balancing Article 8 and 10 claims generally, identified the following relevant factors:

> the context of the comments, the measures applied by the company in order to prevent or remove defamatory comments, the liability of the actual authors of the comments as an alternative to the intermediary's liability, and the consequences of the domestic proceedings for the company.[54]

These cases, however, all concern the traditional, publisher model way of thinking about responsibility for content and behaviours on the internet, where the focus of control is the point in time after the content has been created/posted. This framing arises even when the Court has noted specific characteristics of the internet.

2.5 The Right to be Forgotten

One final group of cases concern the so-called 'right to be forgotten'. The starting point is the EU Court of Justice decision: *Google Spain*, a ruling on the Data Protection Directive (the predecessor to the General Data Protection Regulation (GDPR)). The case arose from a complaint about the fact that when the complainant's name was searched on Google, the results included newspaper reports of more than a decade old about the complainant's then impending bankruptcy. The ECJ upheld the claim for the delinking of those stories. The Court noted the impact of the search engine was to:

> enable[...] any internet user to obtain through the list of results a structured overview of the information relating to that individual that can be found on the internet— information which potentially concerns a vast number of aspects of his private life and which, without the search engine, could not have been interconnected or could have been only with great difficulty— and thereby to establish a more or less detailed profile of him. Furthermore, the effect of the interference with those rights of the data subject is heightened on account of the important role played by the internet and search engines in modern society, which render the information contained in such a list of results ubiquitous.[55]

While the second sentence notes a point already made by the Court of Human Rights (the potential for greater intrusion into privacy and data protection rights), the preliminary argument – about the ability to create structured dossiers – is unique to search engines and similar navigation tools and identifies how they particularly impinged on privacy and data protection rights. Nonetheless, the right to privacy does not take automatic precedence – even though the framing of the directive was to ensure a high level of protection. The Court of Justice noted in particular the right of the public to have access to information, particularly if the claimant played a role in public life.

The Court of Justice has ruled subsequently on the right to be forgotten. In *Manni*,[56] a case which concerned a public register of companies which included information about insolvency, the Court held that the balance in this instance did not come automatically down in favour of deletion – though the question was for the national authorities to decide on the facts. While the court did not dwell on this, by contrast to *Google Spain*, the complaint was against the creator of the information not an indexing intermediary; the register did not have the general range of information nor the ubiquity of the search engine.

More recently, the Court has also held that Google's obligation to de-index did not extend beyond the EU. In its ruling the Court noted that:[57]

- the right to the protection of personal data is not an absolute right, but must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality; and
- the balance between the right to privacy and the protection of personal data, on the one hand, and the freedom of information of internet users, on the other, is likely to vary significantly around the world.

Although Article 17 GDPR has struck a balance between freedom of expression and privacy in relation to the EU internal situation, it is silent as to the legislature's view as to the proper balance externally; the Court did not take that step itself. It did note that it is for the search engine operator to take sufficiently effective measures to prevent or, at the very least, seriously discourage internet users in the Member States from gaining access to the links that had been delisted in the EU using a search conducted on the basis of that data subject's name (para 70). However, while 'EU law does not currently require that the de-referencing granted concern all versions of the search engine in question, it also does not prohibit such a practice' (para 72).

These issues have come before the European Court of Human Rights in the context of Article 8 on the rather particular facts of *ML and WW v Germany*.[58] The applicants had been convicted of murder. During their imprisonment they enlisted the support of the media to try to obtain their release. In 2000, a radio station carried a story about the murder. The full names of the applicants were mentioned. In 2007 and 2008, the applicants were released from prison. A transcript of the report was available on the website of the radio station until at least 2007. ML and WW sought to have the transcript anonymised, and the legal action led to more media coverage. The impugned articles appeared in search engine results but the applicants did not make applications for search engine delisting. The national courts held that the balance between privacy and freedom of expression in this instance lay with freedom of expression. In considering the balance to be made between the conflicting rights, the Court made a distinction between the media and search engines: the media writes the stories and makes the information available; search engines only contributed to the distribution of that information. It seems that the activities of the media then lie at the core of the right to freedom of expression. Against this background, the Court emphasised that the balance of interests between freedom of expression and privacy may lead to different results depending on whether an individual directs her/his request for erasure to a search engine operator or primary publisher (para 97) – and here the case concerned the media not the search engine. Moreover, the Court also reiterated that online communications and their content are more likely than the traditional press to undermine the right to respect for private life and, in this, search engines play an important role (para 91). The Strasbourg court then moved on to balance the interests in the light of its normal case law on the media and privacy (see *Axel Springer*) to find no violation of Article 8. While the result differs from that in *Google Spain*, it seems that the rulings are not inconsistent. Partly the different end-points are due to the different facts (with ML and WW being in a weak position to claim about media intrusion); but, significantly, there is a distinction to be made between the originators of content (particularly the media with its role as watchdog) and vehicles for the dissemination or amplification of that content. This makes express the difference between *Google Spain* and *Manni*, although the point was not discussed by the ECJ. While the Courts do not elaborate on this, there is also a question of the distinction between the delisting of the content (which might still be found by other means) and its removal altogether.[59]

3. Application to the Carnegie Statutory Duty of Care

To assess the regulatory framework in the light of fundamental human rights, it is important to recognise a number of points about the statutory duty of care and human rights intersection with it.

First, it focuses on the processes and steps that the platforms take to try to mitigate harms and may consequently indicate action at a range of points in the chain of communication from content creation through dissemination through to engagement by other users. A number of elements can be identified, for example:

- design features encouraging certain types of behaviour that are not linked to predefined content (e.g. metrics encouraging content designed to be liked);
- editorialising/promotion of content;
- role of navigation and discovery tools, and their impact on content prioritised by the platform;
- tools for users (e.g. mute/blocking tools);
- enforcement of community standards; and
- response to legal orders (e.g. take down orders; blocking orders).

In envisaging that platform operators undertake a risk assessment, the proposal aligns with the Council of Europe recommendation on the roles and responsibilities of internet intermediaries,[60] that:

Internet intermediaries should carry out regular due diligence assessments of their compliance with the responsibility to respect human rights and fundamental freedoms and with their applicable duties. To this end, they should conduct assessments of the direct and indirect human rights impacts of their current and future policies, products and services on users and affected parties, and ensure appropriate follow-up to these assessments by acting upon the findings, and monitoring and evaluating the effectiveness of identified responses.

The UN Special Rapporteur for Freedom of Expression has also noted the need for platforms to engage in due diligence in the context of the introduction of products and rule modifications.[61]

Secondly, there is not a single 'right answer' mandated by the duty of care so platforms will have some degree of choice as to how they approach matters. As noted, this degree of freedom may have an impact on whether a rights claim could realistically be founded, but it is here that the platform's own approach to balancing and protecting human rights comes in.

Thirdly, with the exception of take down mechanisms, these changes/choices are unlikely to be based on particular items of content – they may be said to be neutral as to points of view and as a result potentially less intrusive of speech rights. The communication model envisaged by the Carnegie duty of care can be seen to fall into four stages:

1. creation of content (including access to a platform);
2. dissemination of content (discovery tools and navigation);
3. engagement with content; and
4. deletion of content (moderation and response to complaints and legal actions).

The actions at each stage may engage differently with relevant fundamental freedoms. As regards a basic part of the statutory duty of care – the undertaking of the risk assessment – this in and of itself does not have rights consequences, though decisions made as a result may do so. There is a question here about the development of tools/services/design choices: if a decision is taken to remove a design feature or service, that may trigger a rights-based claim; if there is a decision not to develop/launch a product (for example, because it is too expensive or does not seem to be a good idea for another reason), can we say the same? If we do, does that imply that users have rights for service operators to provide certain features?

3.1 Creation of Content

As noted above, at this stage of jurisprudential development, where a service chooses to restrict its service, it is unlikely that users have a right of access to a particular service (provided refusal does not constitute discrimination); subscription services are likewise not incompatible with users' rights. It may be permissible for a State to require age and/or identity verification in the public interest, at least in certain contexts; one such example would be age verification for porn services. As discussed at 2.2, the court has accepted the need to protect the moral development of children and in *Perrin* placed a low value on obscene speech which was available on the open internet for commercial purposes. These checks might engage Article 8 rights, though it would seem likely that measures against 'fake accounts' (as opposed to those using pseudonyms, or names which reflect the user's identity which might give rise to issues relating to private life and discrimination) would be easier to justify. It is highly questionable whether actions against bots would trigger human rights. Although the Court of Human Rights accepts companies may have rights, they are persons even if legal persons. Bots as yet do not have this status. Serious thought should be addressed before any grant of rights to bots, or even any such rights to those who deploy the bots.[62]

Assuming a user's access to a platform, at this stage of communication we can identify techniques which aid the creation of content and those which limit it. These tools should be understood against the fact that there is a disinhibition effect online,[63] leading users to make decisions that would be less likely in offline environments (e.g. bullying and impulse purchases); a difficult question is whether, if we have been 'nudged' in one direction (especially when it has been to suit the purposes of the person nudging), nudging to counter[64] that is an interference.[65] Of course, the point about nudging is that it leaves open the possibility for the user to resist[66] (though how realistic the chance of such resistance is in practice is open to debate).

The first category falls into two sub-groups: content creation tools provided by the service provider (e.g. furry animal filters or plastic surgery filters) and those that have an indirect impact on content (e.g. reward systems and metrics). While augmented reality (and virtual reality) may have positive uses for mental health,[67] there are concerns that some tools have an adverse effect – for example, those voiced by the British Association of Aesthetic Plastic Surgeons. [68] The latter design elements may affect the user to encourage certain types of content (e.g. outrageous) more likely.

There is no case law on limiting tools of creation (e.g. paint, video cameras and such like), though it is at least arguable that State control over access to these tools could be seen as an interference (particularly if no alternatives were to be available). Means of and tools for distribution (e.g. satellite infrastructure) do seem to be covered, so clearly protection is not limited to expression itself. An assessment of intrusion would be fact specific. Imagine a prohibition on micro-cameras in the interests of privacy; is that justifiable, especially if other forms of camera remained available? If deepfake tools (that is the use of

deep learning/AI to create fake videos) were held to be problematic, a prohibition on their use could constitute an interference with freedom of expression, though it might be questioned how significant an intrusion as the restriction might be about means not message, and does not operate as a complete bar to communicating the message somehow. As noted above, in a number of cases the Court has emphasised the importance of ethics, and of fact-checking and reliable information – deepfakes by their nature manipulate the evidence. Assuming this to be attributable to the State, such a ban might be capable of justification in the interests of truthful information, the intellectual property rights of the owners of the original material, or (in the case of deepfake pornography, the dignitarian rights of others). The second category of possible intrusion is removal of metrics or changing incentives (clickbait and the advertising incentives). It is hard to see that this sort of change actually constitutes an intrusion into speech at all.

In terms of limits on speech, at this stage the most obvious are upload filters (operated by the platform rather than something in the hands of the user) and moderation. Both could constitute an intrusion (and possibly not just in relation to Article 10 ECHR but also Article 8) where their deployment could be linked back to a State requirement. The extent to which such intrusions could be justifiable would be fact specific. EU case law makes clear that general monitoring is not permissible taking into account Article 15 e-Commerce Directive, but also the right to privacy (under the EU Charter).[69] In *SABAM v Netlog*[70], the Court of Justice held that the requirement on a social media platform for proactive filtering of content stored by all its users without distinction indefinitely was incompatible with the e-Commerce Directive (as well as the Information Society Directive and the Enforcement Directive), read in conjunction with fundamental rights.

More recently, following the ruling in *Eva Glawischnig-Piesczek v Facebook Ireland Limited*,[71] it is clear that the (automated) searching for multiple instances of the same content which contravenes civil law (or very similar content) is in the view of the Court of Justice acceptable; the Convention court has not ruled on this issue. In terms of assessing the intensity of an intrusion, much may depend on the technology used. An automated system which identifies matches to a database of content already identified as illegal (e.g. child pornography images), which does not return the images themselves, does not analyse data that is not an image match, and does not allow the operator access to the underlying data is far less intrusive than a system which analyses/categorises all content and allows the operator access to the content itself. In this, the objective served would also be an important factor: the prevention of the dissemination of child pornography justifies more than, for example, a system which tries to stop users posting any 'rude words' at all. It should, however, be remembered that take down, which operates *ex post* and which may have some legal process attached to it to determine the appropriateness of the take down request, is less severe than upload filters, which operate *ex ante*. Where the automated removal of content that is exactly the same may point back to the original processes surrounding the deletion of the original material as safeguards for human rights considerations, this becomes less tenable as a position the further from an exact match we travel. The ECJ in *Glawischnig-Piescek* accepted that the material need not be exactly the same, but that it should be sufficiently similar so that the platform operator was not required to make its own decision as to whether the content was acceptable. Currently, a case is pending before the Strasbourg court concerning the application of filters to an 'extremist' music video.[72]

Using AI more generally may be more problematic. As the UN Special Rapporteur on Freedom of Expression has noted in his report on AI, 'algorithms are today not capable of evaluating cultural context, detecting irony or conducting the critical analysis necessary to accurately identify, for example, "extremist"

content or "hate speech"', with the risk of over-blocking, under-blocking or both.[73] There might also be concerns that these tools embed the cultural assumptions of the designers, or reflect the mores of those included in the training data, resulting in discrimination in the sorts of content identified.

Finally, it might be considered that some features (if they can be brought within the rights framework) might constitute an interference with the 'negative freedom' of being silent or not being compelled to communicate; for example, those that automatically notify friends/followers what a user has looked at. Given the Court's broad approach to the scope of freedom of expression, even a feature on users' accounts that automatically notifies other users when a user has posted content could constitute such an infringement if the user has not chosen to share that content with those notified.

3.2 Dissemination of Content

Given the scale of most platforms, the platforms have a significant role in promoting content – gone are the days when platforms provided a feed just chronologically. There is evidence to show that users are highly influenced by the order in which items are presented in terms of choice of which content to engage with; this is even before autoplay features (which essentially change platforms from 'pull services' to 'push services') are deployed. *Wired* reports that YouTube's changes to its algorithm that prioritised science claims which have some evidence behind them over those that do not, have resulted in a reduction in the number of views containing misinformation in the US by 50 percent.[74] 'Frictionless communication' is seemingly an objective for many online businesses to keep users on the site (and providing more data), so as to encourage behaviour that is 'self-defeating' (e.g. watching one more video); there are concerns about the impact on mental health of online activity because it takes up the time that might otherwise be spent on well-being enhancing activities).

An additional concern relates to 'filter bubbles' – that is that personalisation techniques bring to our attention only 'more of the same', a problem which is exacerbated by a human tendency to favour information that confirms that which we already believe or to follow/friend people who share our views. While an individual certainly has a right to receive information he or she wants from a speaker who wishes to speak, the Court has noted that the State is the guarantor of pluralism, which might suggest a positive obligation under Article 10 to ensure a more varied information diet is promoted; this may indeed tie in with our right to form opinions rather than be manipulated by one-sided information. The Council of Europe has noted:

> Fine grained, sub-conscious and personalised levels of algorithmic persuasion may have significant effects on the cognitive autonomy of individuals and their right to form opinions and take independent decisions. These effects remain underexplored but cannot be underestimated.[75]

The Declaration envisages that some action should be taken about 'over-personalisation' and other forms of manipulation. While the UN Special Rapporteur on Freedom of Expression recognised the dangers of data collection on freedom of opinion,[76] the question of whether targeting and personalisation constitute an interference with our rights either to receive information or to hold an opinion has not been considered to any great degree,[77] though it is accepted that one-sided information (specifically from the State) is problematic.

Some platforms have chosen to make value judgments about content. Fact-checking has been introduced, particularly in relation to concerns about fake news; some groups suggest that 'corrections' should reach all those who saw the original 'inaccurate' post. There might be concerns about the appropriateness of this both from the characterisation of the problem (e.g. not all content is a question of fact) and the effectiveness of the remedy; but the fact of providing more information about a topic is unlikely to constitute an infringement of freedom of expression of the original speaker (though the content may engage Article 8 or 9 rights), or the right to information of the affected audience. Pinterest has removed anti-vaxx results from its internal search function. In the context of self-harm, it has also banned content that encourages suicide or self-injury. This is within the scope of the platform's own freedom of choice. On top of this negative response, the company introduced a process called 'compassionate search', developed in conjunction with mental health specialists, which aims to give users access to information about and techniques for dealing with a range of mental health issues including self-harm. In this its approach is different from other platforms which seemingly rely on providing information (Google) or getting you to share with a friend (Facebook). The search results are unaffected by this process, which aims to mitigate harms that could be exacerbated (because of the risks of contagion and normalisation) even if not caused by the platform. Providing this support is not a violation of users' rights to free expression or to information.

Requiring the platforms to think about the factors they take into account when prioritising content (and the side effects of that) is unlikely to engage a user's rights; it is trite to say that freedom of speech does not equate to the right to maximum reach (and in this, note the approach of the Court of Human Rights in *NL and WW*, above). Such an obligation may however be viewed from two other perspectives. The first is the right of the platform to choose which content to promote; this is equating the running of a platform with editorial control such as that exercised by a cable operator choosing channels. The second takes into account the right of individuals to receive information; this should be understood against the State's obligation to safeguard pluralism. This obligation has been seen in broadcasting cases, but the reasoning of the court is not limited to that context. In this context, we can distinguish between obligations to counter the so-called 'filter bubble'; obligations to prioritise certain sources (analogous to the prominence requirements for public service broadcasting – in that context, they are actors who are subject to obligations in relation to accuracy, range of voices and impartiality, important if an increasing number of people obtain their news from social media); and actions to focus on 'reliable' sources (this could include content that has been watermarked so source is known or so it is clear that the content has not been amended; or to avoid content that has been fact-checked as inaccurate or manipulated). Again, these are viewpoint neutral sources and the nature of the mechanism is that it does not remove content, thereby mitigating net adverse effects on the right to freedom of expression.

As with artificial intelligence generally, the use of this technology needs to be considered carefully in the context of navigation and discovery not just in relation to recommender tools (set on default autoplay), or search results but also in relation to autocomplete suggestions which nudge users' search choices. There is a risk that these tools could embed cultural assumptions of dominant groups in society even if they do not go as far as portraying negative perceptions of minorities.

Another issue where platforms could consider the risks of their services more is in the context of targeting adverts (whether political, issue-based or for products and services). While some targeting may be beneficial (e.g in terms of relevant products or not sending vulnerable groups inappropriate advertising), it seems clear that the feature may be abused – there have been claims that Facebook's advertising can be

used to discriminate against certain groups[78] in relation, for example, to housing or jobs. While commercial speech may fall within the scope of Article 10, it has a lower level of protection than political speech and justifications may be more easily proposed. Indeed, in the context of discriminatory practices, it may be that a State is under a positive obligation to redress the balance. Restrictions on certain sorts of speech (e.g. tobacco advertising) have been justified and it would seem that insofar as a company's choice not to allow targeting in relation to some issue-based advertising which constitutes misinformation (e.g. anti-vaxx speech) constitutes an infringement of speech rights, it would likewise be capable of justification – here public health is in issue. In any event, targeted advertising (which might also implicate privacy rights of users) is another instance where we might distinguish between the right to speak and the desire to reach the maximum audience possible.

3.3 User Engagement

Many design features of online platforms are designed to be easy to use and capture our attention – and these affect how users engage with content. These features include likes or retweets, which capture and reward attention and prompt frequent rechecking; bottomless pages which have no natural breaks so users continue scrolling down; and swipes and 'streaks' which make browsing feel like a game, and, indeed, features like 'streaks' can pressure users to engage in communication when they might not otherwise feel so inclined. Were this to be required by a state actor, it might constitute an interference with a person's right not to speak, or concerns about propaganda and the manipulation of opinion. Push notifications also encourage engagement with a service and some features encourage a fast response. It is arguable that speed of communication plays into human unwillingness to put cognitive effort into discerning the difference between information, disinformation and misinformation as well as other scams, and might therefore contribute to the spread of 'fake news'.[79] It is hard to see that restricting push notifications would affect the user's freedom of speech, though it could be argued that such a restriction could affect the rights of the platform. In such a case, it seems that automated speech would be an issue and, if counting as speech (and it is arguable that this sort of notification is not a communication from a person), would surely be of a variety that is not particularly highly valued.

Another design choice issue concerns the ease with which content from other sources can be embedded or forwarded. In the context of fake news, for example, there is a concern that the resulting decontextualisation means that recipients (which, in unlimited groups, could constitute a large number) are at greater risk of disinformation/misinformation – especially where the 'group' framing tends to intimacy and therefore belief of what is being sent. Here again, factual differences affect the issue of whether there is an intrusion and the significance of any intrusion. For example, seeking to counter information flows by removing a retweet facility still allows a user to share the information. It is however harder to do because the user may need to go to the 'effort' of copying and pasting or even retyping a link or content. If this is an intrusion, surely it is an insignificant one that could be readily justified? Is slowing people down so that they can better appreciate content, especially as regards onward sharing of content, really a violation of those people's right to formulate their own opinion and express their views, or rather – to the contrary – a mechanism to support those rights?

Finally, user engagement is affected by the tools available to users – specifically those that allow users to control their information flows, so that they are not necessarily flooded with more of the same or can block certain other users or types of content, for example. Instagram has been developing some tools to allow users to block bullies[80] (or others) without having to engage with the content first. The tool comments on

the user's posts from a person that user "restricted" will only be visible to the restricted person, though the user can choose to view the comment; approve the comment so everyone can see it; delete it; or ignore it. No notifications in relation to comments from a restricted account will be received by the user. The blocked person has no speech right to reach people who do not want to hear that person.

Some design proposals may try to slow speech down – for example, asking a user if they really mean a particular post (based on keywords) or reminding users of terms of service. Instagram has introduced a feature which detects when posts might be considered offensive or hurtful and flags this so that users may reflect on whether they want to rephrase their comments.[81] In these situations, speech may still happen. It is debatable the extent to which these design features, which still allow users the choice to go ahead, constitute an intrusion, especially if the design features are countering nudges to speech of which the user may not be aware.

3.4 Complaints, take downs

While the action at earlier stages may reduce some of the problems perceived to flow from the design and use of social media platforms, it seems likely that there will always be a need to respond to concerns about particular items of content: this is the context at which freedom of expression concerns take centre stage. In particular, domestic laws restricting specific categories of speech should be clear about the nature of the offending content, though often these terms are quite broad or vague. This has given rise to concerns, for example, in the context of fake news[82]. Nonetheless, it is important to remember that even if some of the mechanisms responding to a duty of care do constitute an intrusion into freedom of expression, they may be justified (indeed, as noted, some speech may lie outside protection altogether) or required in the protection of other rights (e.g. right to reputation). The Court of Human Rights has accepted that "an act motivated by a personal grievance or a personal antagonism or the expectation of personal advantage, including pecuniary gain, would not justify a particularly strong level of protection".[83] The English courts have also noted that where speech rights are 'misused' (e.g. blackmail, harassment, misuse of private information), they will not carry much weight.

When considering items of content, it is important to recognise that there is a difference between complaints based on community standards which are set by the platforms and which the platforms are legitimately entitled to interpret (though this must be within the limits set by the law as from time to time understood); and the application of general law, whether criminal, civil or administrative. Further, as the Council of Europe has noted, interference with the flow of information may take place in accordance with the platform's own policies as well as in accordance with the law.[84] As regards community standards in addition to the rights of users we should consider the rights of the platform. In *Lee v Ashers*,[85] the Supreme Court had to decide a discrimination case in which a baker had refused to make a cake in support of gay marriage on religious grounds. Lee had ordered the cake to be decorated with a message: "Support Gay Marriage", to mark the International Day Against Homophobia and Transphobia. The Supreme Court reversed the decision of the Court of Appeal, ruling that Lee was not discriminated against. In this case, the Court found for the baker because the objection was not to Lee himself but to the message. Further, the Court noted Article 10 also involves the right not to speak. Decorating the cake with the message would infringe that right. This sort of reasoning could apply to platforms too, which may have speech rights (or in the context of the EU Charter rights to run a business).

As regards community standards, we would expect platforms to have an effective mechanism for responding to user complaints, and it should be this system that is considered for a duty of care. The independent regulator should not be taking decisions in individual cases. Rather its role would be overseeing the operation of the complaints and take-down processes. In *CM/Rec* (2018) 2[86], the Council of Europe suggests that "intermediaries should take reasonable and proportionate measures to ensure that their terms of service agreements, community standards and codes of ethics are applied and enforced consistently and in compliance with applicable procedural safeguards" (para 2.1.5).

This obligation does not bite directly on freedom of expression (a platform's desire not to take responsibility is not an example of freedom of expression) though it may have consequential effects on users' rights. In this it is important to remember that the rights in issue might not be limited to the expressive right of the person whose post is in issue, but may include the expressive rights of others as well as other rights (e.g. safety, privacy, religion, association). Process obligations as to fair and effective disposal of such complaints as have been recognised elsewhere are an appropriate mechanism, but its processes should not be set up so that they automatically favour one right over another, or the perpetrator over the complainant, or one group of people over another.[87] So, rights to appeal or contest should in principle be available to all sides in a dispute.

Some platforms (e.g. Instagram) have introduced policies[88] whereby users who have a certain amount of content that violates community standards, or a certain frequency of violations within a set time frame may be banned from the platform. Should this sort of decision be attributable to the State (which as noted, is not necessarily the case), then there might be questions about proportionality depending on the sorts of violations in issue, as well as the quantity threshold. It is important to remember that the European Court of Human Rights has accepted that "it is legitimate to try to ensure a minimum degree of moderation and propriety and that a clear distinction must be made between criticism and insult" – and this is so even in the context of political struggles and even concerning the reputation of a controversial politician.[89] Of course, there is a difficult line to draw between this and the vulgarity found acceptable in *Tamiz*.[90] Further, while such a ban might have severe consequences on the banned user's rights (not just freedom of expression, but possibly the right to receive may be Articles 8 and 11 ECHR), it is a less severe restriction than the collateral censorship in *Akdeniz*.

4. Conclusions

It is difficult to analyse the strength of a rights-based claim of a user because rights in a formal sense bite on States not private actors. While States may be under positive obligations to intervene to protect users' rights, they must achieve a fair balance between conflicting rights. In this, it is significant that the platforms may hold rights to be taken into account too, as well as the rights of other users. Even accepting that the framework imposing a duty of care may lead to some responsibility on the part of the State, it seems that the duty of care could be justified. While the different actions might have a greater or lesser effect on rights, all could be acceptable from a human rights perspective depending on the context (the size of the provider; the nature of the service; the objective aimed at especially bearing in mind States' positive obligations) and on the safeguards provided.

Professor Lorna Woods
University of Essex

Endnotes

1 See the Carnegie UK work at: https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media: in particular "Online Harm Reduction: a Duty of Care and a Regulator" (April 2019): https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

2 For a discussion of the changing approach to position of corporations and the distinction between 'hard' and 'soft' law see e.g. B. Choudhoury 'Balancing soft and hard law for business and human rights' (2018) 67 ICLQ 67

3 Report of the Special Representative of the Secretary General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie – Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy Framework" (A/HRC/17/31), 21 March 2011.

4 Section 6 Human Rights Act

5 Section 149 Equality Act

6 *Re S (A Child) (Identification and Restrictions on Publication*) [2004] UKHL 47)

7 *PJS* [2016] UKSC 26

8 Ibid, para 20. The EU also recognises the need to ensure a fair balance in which not right has automatic priority: Coty Germany, C-580/13, EU:C:2015:485, para 34

9 *Appleby v UK* (44306/98), judgment 6 May 2003

10 K. Tiffany, Inside R/Relationships, the Unbearably Human Corner of Reddit, *The Atlantic*, 23 October, 2019, available: https://www.theatlantic.com/technology/archive/2019/10/reddit-moderation-relationships-subreddit-memes/600322/.

11 Code of Conduct available: https://mew.toot.cat/mw/Pub/toot.cat/CoC [accessed 2 December 2019]

12 Available here: https://mastodon.org.uk/about/more [accessed 2 December 2019].

13 *Stomakhin v Russia* (52273/07), judgment 9 May 2018

14 *Lingens v Austria* Series A/103, judgment 8 July 1986, para 41; *Observer and Guardian v UK* (13585/88), judgment 26 November 1991; *Jersild v Denmark* (15890/89), judgment 23 September 1994.

15 *Handyside v UK* Series A/24, judgment 7 December 1976, para 49

16 *Women on Waves v Portugal* (31276/05), judgment 3 February 2009

17 See e.g. *Norwood v UK* (23131/03) decision 16 November 2004

18 *Sürek v. Turkey (no. 1)* (26682/95) ECHR 1999-IV, para 62,

19 *Medžlis Islamske Zajednice Brčko* (17224/11), judgment 27 June 2017 [GC], para 108

20 *ES v Austria* (38450/12), judgment 25 October 2018, para 53

21 *Pastors v Germany* (55225/14), judgment 3 October 2019

22 *Eon v. France* (26118/10), judgment 14 March 2013, paras 60-61.

23 *Nikowitz and Verlagsgruppe News v Austria* (5266/03), judgment 22 February 2007, para 25.

24 *Sousa Goucha v Portugal* (70434/12), judgment 22 March 2016, para 51

25 *Sinkova v Ukraine* (39496/11), judgment 27 February 2018

26 *M'Bala M'Bala v France* (25239/13) decision 20 October 2015

27 *ES v Austria* (38450/12), judgment 25 October 2018

28 Ibid, para 43

29 Ibid, para 44

30 *Belkacem v Belgium* (34367/14), decision 27 June 2017

31 For a discussion of some of the issues in this area see e.g. Kuhn, "Reforming the approach to racial and religious hate speech under article 10 of the European Convention on Human Rights' (2019) HRLRev 119-147;

32 *Axel Springer v Germany* (39954/08), judgment 7 February 2012 [GC]

33 This approach has also been adopted by the domestic courts: *PJS v News Group Newspapers* [2016] UKSC 26

34 *Bensaid v UK* (44599/98), judgment 6 February 2001, para 47

35 *Perinçek v Switzerland* (27510/08), judgment 15 October 2015 [GC].

36 *Axel Springer v Germany* (39954/08), judgment 7 February 2012 [GC]

37 *Söderman v. Sweden* (5786/08), judgment 12 November 2013

38 *MC v Bulgaria* (39272/98), judgment 4 December 2003

39 *KU v Finland* (2872/02), judgment 2 December 2008.

40 *Volodina v Russia* (40419/19), communicated.

41 *Durdevic v Croatia* (52442/09), judgment 19 July 2011

42 *Király and Dömötör v. Hungary* (10851/13), judgment 17 January 2017, para 80.

43 *Tamiz v UK(3877/14)* decision 19 September 2017

44 *Burlya v Ukraine* (3289/10), judgment 6 November 2018, paras 167-170

45 *Opuz v Turkey* (33401/02), judgment 9 June 2009.

46 UN Special Rapporteur on Violence Against Women Report on online violence against women and girls from a human rights perspective (A/HRC/38/47), paras 12 and 30.

47 e.g House of Commons and House of Lords Report of the Joint Committee on Human Rights on Democracy, para 14 and 49

48 House of Commons and House of Lords, Joint Committee on Human Rights, The Right to Privacy (Article 8) and the Digital Revolution (HC122/HL Paper 14), 3 November 2019

49 *Perrin v UK* (5446/03), decision 18 October 2005.

50 *Einarsson v Iceland* (24703/15), judgment 7 November 2017.

51 see e.g *Bartnik* v. *Poland* (53628/10), decision 11 March 2014; *Pihl v Sweden* (74742/14), decision, 7 February 2017, paras 26-27; *Høiness v Norway* (43624), judgment 19 March 2019, paras 65-67

52 *Niskasaari and Otavamedia Oy v. Finland* (32297/10), 23 June 2015, paras 9 and 54-59.

53 *Yildirim v. Turkey* (3111/10), judgment 18 December 2012

54 *Delfi v Estonia* (64569/09), judgment 16 June 2015 [GC], paras 142-3; *MTE v Hungary* (22947/13), judgment 2 February 2016, para 69; *Høiness v Norway* (43624), judgment 19 March 2019, para 67

55 *Google Spain, and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* (Case C-131/12) judgment 13 May 2014 [GC], ECLI:EU:C:2014:317, para 80.

56 *Camera di Commercio, Industria, Artigianato e Agricoltura di Lecce v Salvatore Manni* (Case C-398/15), judgment 9 March 2017, ECLI:EU:C:2017:197.

57 *Google LLC v CNIL and ors* (Case C-507/17), judgment 24 September 2019, para 60.

58 *ML and WW v Germany* (60798/10 and 65599/10), judgment 28 June 2016

59 See also *Wegrzynowski and Smolczewskiv. Poland* (33846/07), judgment 16 July 2013, in which the Court held that the press cannot be expected to remove all traces of a defamatory publication from its archives.

60 Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, 7 March 2018, , para 2.1.4, available: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 [accessed 3 December 2019].

61 Report of the Special Rapporteur on Freedom of Expresion (A/HRC/38/35) para 55

62 For a discussion of some of the issues, albeit from a US perspective see F Pasquale, 'Preventing a Posthuman Law of Freedom of Expression' 26 February 2018, available: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 [accessed 3 December 2019]. More generally about robot rights, see concerns expressed in an open letter, available here: http://www.robotics-openletter.eu/

63 E. Costa and D Halpern, The behavioural science of online harm and manipulation, and what to do about it: An exploratory paper to spark ideas and debate (Behavioural Insights Team, 2019), available: https://www.bi.team/wp-content/uploads/2019/04/BIT_The-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it_Single.pdf [accessed 3 December 2019):

64 N. Levy 'Nudges in a post-truth world' 43(8) Journal of Medical Ethics, DOI http://dx.doi.org/10.1136/medethics-2017-104153

65 On ethics of Government nudging, see e.g. Sunstein, 'The Ethics of Nudging' (2015) 32(2) *Yale Journal on Regulation* 413

66 Sunstein and Thaler 'Libertarian Paternalism is Not an Oxymoron' (2003) Chicago Unbound: Public Law and Legal Theory Working Papers, available: https://chicagounbound.uchicago.edu/public_law_and_legal_theory [accessed 3 December 2019]

67 Riva et al 'Transforming experience: The potential of augmented reality and virtual reality for enhancing personal and clinical chance' (2016) Frontiers in Psychiatry 164, DOI https://dx.doi.org/10.3389%2Ffpsyt.2016.00164 [accessed 3 December 2019]

68 https://www.telegraph.co.uk/news/2019/03/26/snapchat-instagram-filters-should-carry-health-warning-teenagers/

69 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL* (SABAM) (Case C-70/10), judgment 24 November 2011, ECLI:EU:C:2011:771

70 *SABAM v Netlog* (Case C-360/10), judgment 16 February 2012, ECLI:EU:C:2012:85

71 *EvaGlawischnig-Piesczekv Facebook Ireland Limited* (Case C-18/18), judgment 3 October, 2019, ECLI:EU:C:2019:821

72 *Alekhina and Others v. Russia* (38004/12), communicated

73 Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression (A/73/348), 29 August 2018.

74 E Grey Ellis, The Influencer Scientists Debunking Online Misinformation Wired 13 November 2019, available: https://www.wired.com/story/youtube-misinformation-scientists/ [accessed 3 December 2019].

75 Declaration on the manipulative capabilities of algorithmic processes (Decl (13/02/2019)1), para 9, available: https://search.coe.int/cm/pages/result_details.aspx?objectid=090000168092dd4b [accessed 3 December 2019].

76 Report of the Special Rapporteur on Freedom of Expression (A.HRC.29.32), paras 19-21

77 S. Alegre 'Rethinking Freedom of Thought for the 21st Century' [2017] EHRLR 221

78 Speicher et al Potential for Discrimination in Online Targeted Advertising (2018) 81 Proceedings of Machine Learning Research 1, available: http://proceedings.mlr.press/v81/speicher18a/speicher18a.pdf [accessed 4 December 2019].

79 Some research suggests that fake news spreads further and faster than facts: Soroush Vosoughi, Deb Roy and Sinan Aral, "The spread of true and false news online" (2018) 359 Science 1146-1151

80 Instagram, Empowering our Community, available: https://instagram-press.com/blog/2019/10/02/empowering-our-community-to-stand-up-to-bullying/ [accessed 4 December 2019]

81 Instagram, Our Commitment to Lead the Fight against Online Bullying, available: https://instagram-press.com/blog/2019/07/08/our-commitment-to-lead-the-fight-against-online-bullying/ [accessed 4 December 2019]

82 OSCE: "Joint declaration on freedom of expression and "fake news", disinformation and propaganda", 3 March 2017, available: https://www.osce.org/fom/302796

83 *Kudeshkina v Russia* (29492/05) judgment 26 Febriary 2009, para 95

84 Recommendation CM/Rec(2018) 2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, 7 March 2018 para 2.1.3, available: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 [accessed 3 December 2019]

85 *Lee (Respondent) v Ashers Baking Company Ltd and others (Appellants) (Northern Ireland)* [2018] UKSC 49

86 Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, available at: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14

87 There are concerns as to whether equal access and equal treatment happens: R. van Loo 'The Corporation as Courthouse' (2016) 33 *Yale J Reg* 547

88 Instagram, Changes to Our Account Disable Policy, 18 July 2019, available: https://instagram-press.com/blog/2019/07/18/changes-to-our-account-disable-policy/ [accessed 3 December 2019]

89 See *Genner v Austria (*55495/08), judgment 12 January 2016, para 36, reiterated in e.g. *Annen v Germany (No 6) (*3779/11), judgment 18 October 2018, para 24

90 *Tamiz v UK(3877/14)* decision 19 September 2017