"Doing More" to End Sexting - Facts, fictions and challenges in the policy debate on young people's sexting behaviour

Andy Phippen (Plymouth Universit)y, Margaret Brennan (University College Cork) and Jennifer Agate (Senior Associate, Foot Anstey)

In November 2016, Jeremy Hunt, the UK Health Secretary, gave evidence to the Commons Health Committee's investigation into suicide prevention, which raised concerns about the impact of technologically facilitated social behaviours on children's mental health and wellbeing. In his submission, Mr Hunt suggested that in order to address issues that young people face, particularly sexting, online service providers should be providing technological intervention that would prohibit such practices occurring. While "sexting" has been variously defined academic and practice circles, for the purposes of this article we will use the definition of the National Society for the Prevention of Cruelty to Children:

"Sexting is when someone shares sexual, naked or semi-naked images or videos of themselves or others, or sends sexually explicit messages." (https://www.nspcc.org.uk/preventing-abuse/keeping-children-safe/sexting/)

In this article we explore the nature of the proposal presented by a senior politician around this important child safety and wellbeing issue and argue that it follows a flawed, yet reoccurring claim that it is the role of online service providers, not other stakeholders, to provide preventive solutions to social problems. In this exploration we submit that such reliance on technology, without parallel concentrations of effort in the development of other, more human, interventions, will give rise to "solutions" that will be, at best, limited in efficacy, and at worst, doomed to fail at the expense of children's rights and wellbeing.

The influence of digital technology

During his evidence, Mr Hunt raised the issue of the influence of digital technology on children's mental health and wellbeing, which is certainly to be applauded. However, Mr Hunt's evidence suggested that this issue could, and should, be resolved by online service providers – thus adding to a growing list of Government demands for service providers to tackle all manner of social problems that are mediated via their platforms. These include, for example, the identification of hate speech, the preclusion of sexual grooming, the control of access to "inappropriate" content, the identification of cyberbullying and the control of "offensive" posts (http://www.huffingtonpost.co.uk/entry/twitter-facebook-face-statutory-code-of-practice-online-abuse-digital-economy-bill-baroness-jones_uk_58931e4de4b0302a153d2df4). All of these issues may be addressed to some degree through technical intervention, but before we move to consider such interventions as "solutions", it would be sage to recall Ranum's law (http://www.kianmeng.org/2014/02/ranums-law.html), named after the renowned IT security specialist Marcus Ranum:

"You can't solve social problems with software."

While these issues may arise across, for example, social media and mobile platforms, these are problems that are manifesting via digital technology, rather than being unique to it. While the impact, reach and persistence of these problems may differ when mediated though digital

platforms, issues such as grooming, bullying and hate speech are nothing new, having existed in the social environment long before the advent of social media.

However, notwithstanding the broad set of stakeholders in mental health and suicide prevention, at least for this evidence session, the Health Secretary reasoned that if a problem is facilitated via technology, it should be prevented by technology, and furthermore, that service providers could develop software to detect and prevent harmful behaviours or identify distress. Mr Hunt focussed on two issues in particular – the distribution of self-generated sexually explicit images (a form of "sexting") and cyberbullying:

"I just ask myself the simple question as to why it is that you can't prevent the texting of sexually explicit images by people under the age of 18, if that's a lock that parents choose to put on a mobile phone contract. Because there is technology that can identify sexually explicit pictures and prevent it being transmitted…

…I ask myself why we can't identify cyberbullying when it happens on social media platforms by word pattern recognition, and then prevent it happening. I think there are a lot of things where social media companies could put options in their software that could reduce the risks associated with social media, and I do think that is something which they should actively pursue in a way that hasn't happened to date."

Policy directions

Mr Hunt's view follows recent policy directions around online child safety and protection – that technology providers should "do more" to stop harmful and illegal activities involving children their platforms and services. In his evidence Mr Hunt outlined several, apparently straightforward, proposals to address the issues of sexting and cyberbullying. While the focus of this article is on sexting, it is first worth exploring, in passing, the feasibility of Mr Hunt's suggestion that cyberbullying may be prevented through "word pattern recognition" as this also has implications for sexting; cyberbullying may be sexualised in its presentation and therefore these issues may not be readily divorced from each other.

There is sometimes a view that because cyberbullying might feature specific words (for example "hate", "kill", "hurt", "beat up"), in specific contexts, it should be easy for an algorithm to identify them. However, the problem with this position is that while contemporary algorithmic approaches are highly effective for keyword and pattern recognition tasks (for example, identifying the proximity of certain words to others), they are not particularly effective in making contextual interpretations or semantic inferences. For example, if someone were to post "I hate you" to a contact on a social media site, the supporting sentiment may well be harmful; they may be attacking the contact and wishing for them to feel upset as a result of these words. In this same instance however, the supporting sentiment may be wildly different; the poster may be attempting to say something amusing about their contact's profile picture. This situation is further complicated by the fact that establishing the child's own perception of his/her victimization is important in identifying bullying incidents and responding appropriately. How, then, based on the limited behavioural information available in the online exchange, would we expect a social media company to confirm a case of cyberbullying and intervene appropriately? At what threshold should any social media company intervene in children's communications?

Technologically mediated intervention

These challenges become particularly sensitive where issues of mental health and suicide are concerned. If we are to believe the Health Secretary's perspective on these matters, it is feasible to expect software to identify these problems in online spaces and to support appropriate intervention. While some technologically mediated intervention already exists, it is low key, unobtrusive and often keyword-linked (for example, carrying out a Google search for "suicide" will result in links to the Samaritans). The sorts of expectations around technological intervention espoused by Mr Hunt risk a situation where online service providers' responses become intrusive, particularly given the private sector nature of most social media providers. Should we really be expecting privately owned companies with little public accountability to be making judgements on the state of users' mental health based upon their online behaviour, and to intervene if their algorithms suggest they are at risk?

A Google search for "I want to kill myself" on the popular blogging site Tumblr results in over 29,000 responses across its pages and posts. Should we expect this service provider to assess and potentially intervene in all of these cases? And if so, *how* would this be done? While there is clearly an element of posting on Tumblr that does relate to self-harm and suicide, should it be down to Tumblr solely to determine what to do about these posts? Mental health and suicide prevention are complex, nuanced and idiosyncratic issues that command sensitive and multi-faceted approaches to detection and intervention; approaches significantly more complex than simple keyword searching, algorithmic interpretation or indeed technology providers can offer.

The issue around detection and intervention becomes even more complex where images and other media feature in the exchange, as arises in the context of sexting behaviour. The premise remains the same – the expectation is that an algorithm can successfully identify a behaviour and an associated sentiment (e.g. that a young person has self-generated an indecent image of themselves and is going to send it to someone) and trigger a series of processes that will prevent this from happening.

What is the threshold for "indecency"?

From a rights and legal perspective this is extremely problematic. Let us first take human interpretation of this problem – could we, as humans, make consistent, accurate decisions on whether or not a self-generated image is indecent? What is the threshold for "indecency"? In legal terms, the Coroners and Justice Act 2009 (Section 62) defines a prohibited image of a child as follows:

1. That the image is pornographic;
2. That the image is grossly offensive, disgusting, or otherwise of an obscene character; and
3. That the image focuses solely or principally on a child's genitals or anal region, or portrays a number of defined sexual acts.

However, this definition would not encompass a large proportion the sorts of images young people have described to us as featuring in sexting exchanges. These can include, for example, images of the top half of the body or photographs of the subject in their underwear. Do we therefore need another definition of what an indecent image is within the sexting context? At what threshold would the images that feature in sexting exchanges become so morally offensive that they merit classification as an indecent image? And how would the categorisation of such images be operationalized? Should we set that threshold at full or partial nudity of specific parts of the body? Could a naked image ever not be indecent? Or,

perhaps, images that depict children in their underwear? And if that is the case, should we also include in our threshold swimwear? Should our definition of indecency also include fetishistic images, for example, images of feet, or perhaps those dressed in PVC? And who should be appointed to rule on such matters? Should we have an arbiter of decency whose judgement is final on these decisions, even though there is a risk that perfectly innocent images may be ruled as indecent within these suggested definitions? For example, someone sending their friend a picture of their new shoes.

While legal definitions are effective for the judicial process, are they effective for defining algorithmic parameters for the identification indecent images in the context of sexting exchanges? Mr Hunt claims that service providers might have "*technology that can identify sexually explicit pictures and prevent it being transmitted."* The reality, however, is far more complex.

Image and statistical recognition

Image recognition algorithms, as applied in the policing of online child sexual exploitation, generally work in one of two ways. Firstly, through some form of hashing – applying a mathematical process to the data in the file to generate a unique value that signifies that specific image. This algorithm might be applied to, for example, the colour of each pixel in the image, the different colours in the image, keywords associated with the images, or image filenames, often referred to as "feature extraction" (Rui et al., 1999). The resulting hash value is an alphanumeric, hexadecimal sequence (typically 32 characters in length), analogous to a digital fingerprint that uniquely identifies that image. This hash value can then be used to identify reoccurrences of that image in other locations, when it is stored with different filenames, for example. The unique hash value is based upon the image composition, or the physical properties of the image, and some, such as Microsoft's PhotoDNA, can accommodate slight changes to the image (for example, changes in colour, making the image black and white, or changing its size).

Statistical "recognition" is a different approach, which, for the purposes of detecting indecent images of children, applies mathematical techniques to identify common characteristics, usually in a set of images. Again this approach is based upon the data in the image (for example, the proximity of different colour pixels to each other, the frequency of images of different colours, specific colours in specific parts of the image, etc.). In more recent times such artificial intelligence (AI) techniques have been applied as an alternative approach to identify common characteristics within candidate child exploitation images, entered into a detection system. However, these techniques do little to understand the meaning of the image; rather they identify similar characteristics within the data held in the image.

In the context of online child exploitation, perhaps the most famous hashing technique is Microsoft's PhotoDNA, which is used within Microsoft's own search services as well as social media platforms and law enforcement initiatives. Its use is extremely successful when it is used within the parameters of its functionality – it is a technique to uniquely identify an image based upon the data within it. Therefore, when other instances of the image appear and the hashing algorithm is applied, the same unique value will be generated and a match identified.

The technique has been very successfully applied for the purposes of image recognition where an image is known, and has already been hashed. However, when we consider the

application of such techniques in determining whether a new, self-generated, image is indecent, there is little in the functionality to help. There is nothing in this technology that currently supports the identification of new indecent images, or makes inferences of indecency in relation to new, previously unseen content. The algorithm can merely confirm whether the image is similar to a known image, previously identified as indecent through human intervention.

With statistical and AI-based techniques in this domain, the "image recognition" component still resides in comparing one set of data with another to identify similarity, the dataset in this case most often being an image, though filename classification approaches may also be used to complement this process (e.g. Peersman, Rashid, Schulze, Brennan & Fischer, 2016). Again, there is no inference made by the algorithm concerning the meaning of the candidate image, just whether the image is a statistical relative of other images that have had similar techniques applied to them. Generally, machine-learning algorithmic approaches rely on being "trained" – being fed collections of pre-classified images (in the case of sexual or indecent images, the images would have already been classified by human intervention as such) to look for data similarities and make predictions, so that when an unknown image is submitted to the learning system, it can make an estimate of the similarity of the image to the training set. However, this is an estimate based on similarity, of data in the files, not a similarity of the meaning of the image. Two indecent images may appear very different in their composition, and whereas we, as humans, will recognise both as indecent (within our own interpretation of what indecency is), current algorithms are unable to make such an inference based on an assessment of the physical properties of an image alone. Therefore, while such systems could be trained with a set of self-generated images, and there might be a likelihood that statistically similar images will be found, this does not mean that such a system would be able to identify all self-generated images, just those that have similarity to the training sets.

Given that most image recognition algorithms "recognise" based upon the physical properties of the image, rather than what the image represents in semantic terms (which is, of itself, extremely difficult to do algorithmically for reasons discussed above), and largely rely on training sets to recognise comparable images, the identification of unknown images, which we might argue a new, self-generated indecent image always would be, is far more complex. Even the most advanced algorithms in this field have acceptance rates that would struggle to be termed reliable without some form in human intervention. In a recent blog post on Algorithmia (http://blog.algorithmia.com/improving-nudity-detection-nsfw-image-recognition/) around an advanced "nudity detection" algorithm (note the significance here that this was nudity, rather than indecency) the testers returned a maximum positive accuracy of 83.64% and maximum negative accuracy of 87%. Therefore, in over 1 in 10 cases, the algorithm would not be able to accurately detect whether an image contained naked skin, let alone whether the image was indecent in nature. In order for such techniques to scale to meet the challenge of sexting, we require levels of accuracy much greater than many state of the art detection algorithms allow. While more recent artificial intelligence tools for the identification of "new", previously unseen child exploitation media files report accuracy rates higher than this for image and video detection (see Peersman et al., 2016), such approaches have not been applied to the task of detecting newly produced, self-generated sexual content depicting children.

Perhaps more concerning is that the perception that algorithms can accurately detect indecent images – something reinforced when the Health Secretary makes such claims in a select

committee hearing – is creating a marketplace for "safeguarding" software that aims to reassure parents that their children are prevented from engaging in the exchange of indecent, self-generated, images. While some "simply" send any image sent from the child's phone to the parent's phone too, others claim advanced image recognition capabilities, and the ability to reliably identify everything from indecency to self-harm. Several such products offer a range of functionalities, such as alerting the parent when such an image is generated, forwarding the image to the parent, or posting a warning on the child's phone about a suspect image.

Given the limitations to the accuracy of image recognition algorithms described above, we would be concerned about the rights and wellbeing of children and families if algorithms were applied as the "solution" to this issue, as called for by Mr Hunt. Should we be happy, even with the rates of sensitivity and specificity maintained by more advanced nudity detection algorithms cited above, that over 1 in 10 children could be falsely implicated in the self-generation of nude images, and that the result of this "detection" could be an alert to their parents? A false positive could result in intra-familial distress, conflict and the kinds of sanctions children so often report they fear, for example, a parent demanding access to, or confiscating their phone when no such sanctions were necessary. With software that simply forwards all images onto the parent's phone without the knowledge or consent of the child, surely this could be framed as a severe challenge to the child's to privacy (one of the fundamental articles of the UN Convention of the Rights of the Child, ratified by most members of the United Nations)?

That is not to say that all algorithmic approaches to image detection are ineffective and that service providers are doing as much as they can. Given that PhotoDNA can generate a unique hash value for images run against the algorithm, once an image has been hashed, there is no reason why the service provider could not identify the image if it was reposted. For example, if a disgruntled partner posted an indecent image of their significant other and that person then requested that image be taken down, once hashed, the service provider should be able to detect reposts and facilitate removal. This is very much the argument in an ongoing legal case in Belfast, where a 14-year-old girl is suing Facebook for failing to identify a reposted image of her even though it had previously been taken down. However, we cannot simply say that service providers need to provide the solutions to these issues, something that certainly seems to be the view of Mr Hunt is we are to return to his evidence at the Health Select Committee. What is clear from what he says is that he believes it is the role of service providers to tackle children's mental health issues through technical intervention:

"I think social media companies need to step up to the plate and show us how they can be the solution to the issue of mental ill health amongst teenagers, and not the cause of the problem. There is a lot of evidence that the technology industry, if they put their mind to it, can do really smart things."

Prevention rather than prohibition

While these technologies can clearly "do smart things" the they work with a precise logic and within the boundaries of processing power that is sometimes challenged by an area such as sexting, which can be a complex, imprecise and subjective concept. Claiming that unreliable algorithms are the solution to a problem that can have a significant impact upon a child's mental health is at best naïve and at worst irresponsible. Surely a more useful approach for government would be one around policy that looks at prevention rather than prohibition?

Prohibitive approaches to social problems have a rich history of failure, and there is little evidence to suggest that such approaches would deliver better outcomes where an issue such as sexting is concerned. The rationale for this contention is that while sexting is facilitated by technology, its roots often lie in non-technological psychosocial challenges related to a lack of respect, consent, self-esteem and awareness in the context of young people's sexual behaviours online (Ringrose et. al. 2012).

While the role of legislation in this space has been well discussed (for example Phippen 2017) the legal position around sexting is, of itself, complex. While under the letter of the law (specifically the Protection of Children Act 1978) the generation and distribution of an indecent image of a child is a criminal offence, exponential increases in the production and exchange of self-generated sexual content have introduced much complexity to this issue, particularly in relation to the public interest test. At the time of the inception of this law it was probably inconceivable that the child subject of the image might also be the creator of the image – this was a law that aimed to protect children from predatory behaviour, it did not conceive of a time where a child might voluntarily chose to self-generate an image on their own device and distribute it to others.

Drawing from our own empirical work with children and young people this is certainly the case. While the volume of self-generation among teens is difficult to measure (for a number of reasons, for example asking children if they have engaged in illegal activity presents ethical and legal challenges for the researcher), it is clear from our discussions with children that this is something that occurs regularly within peer groups, and the age of those self-generating is reducing (Phippen 2017).

As such the application of the law has increasingly created problems with both schools and law enforcement. In the event of any self-generation and exchange of indecent imagery between peers which is brought to the attention of the school safeguarding team, the legally correct response is to report this to the police given that illegal activity has taken place. However, this has the potential to cause a chain of events that result in both the sender and receiver of the image ending up with criminal records. While Crown Prosecution Service guidance (http://www.cps.gov.uk/legal/a_to_c/communications_sent_via_social_media/) in recent times has stated that it is rarely in the public interest to prosecute a minor in such an instance (unless there is predatory or coercive behaviour involved), there have been cases reported where children have ended up with police cautions for such behaviour where evidence of coercion or predation was lacking (for example http://www.telegraph.co.uk/news/uknews/crime/11840985/Teenage-boy-added-to-police-database-for-sexting.html).

More recently the Policing College has issued its own guidance, which allows a sexting incident to be reported and recorded, without the child ending up with a criminal record (http://www.college.police.uk/News/College-news/Documents/Police_action_in_response_to_sexting_-_briefing_(003).pdf). In order to provide a middle ground between the incident going unreported, and the image producer/victim ending up with a criminal record, guidance was issued on something referred to as an "Outcome 21" response:

"Further investigation, resulting from the crime report, which could provide evidence sufficient to support formal action being taken against the suspect is not in the public interest – police decision."

Nevertheless, there is still complexity within this given that all sexting incidents are not the same. While a peer to peer exchange might be consensual, other factors, such as exploitation, coercion, or deception can prompt young people's sexting behaviours. A minor may be coerced into self-generating an image as a result of inter-personal pressure or more malicious activity such as blackmail, which often features threats to redistribute other sexual images of the young person. In cases such as these, there is a public interest in sanctioning the behaviour of the offending party. To its credit, the Policing College advice is cognisant of this, and provides guidance to officers dealing with such situations, for example, distinguishing between redistribution "without malicious intent" and redistribution "with malicious intent with evidence of grooming or coercion". However, what the advice does not provide support with is the interpretation of malicious intent or the identification of coercive practices; these determinations are ultimately made at the discretion of the inquiring officer.

Educational responses

While the legislative process is improving, and there is greater protection for victims as the criminal justice process adapts to this emerging phenomenon with a greater level and understanding and pragmatism it will, again, never be a solution, just as the legality of substance abuse does not prevent engagement with such practices. Ultimately, educational responses are far more likely to provide positive outcomes than either technology or law. This is certainly the view of many young people we speak to, who frequently comment that they never get a chance to talk about issues around sexting in their schools and certainly never get opportunity to ask questions about it. How can we hope that legal protection for victims will work when they are not aware of the protection the law offers them or even that they are victims of a crime. What was clear from the work coming out of the recent Women and Equalities Committee Report on sexual harassment and sexual violence in schools (https://www.publications.parliament.uk/pa/cm201617/cmselect/cmwomeq/91/91.pdf) was that a lot of victims of such crimes are not aware that what is happening to them in unacceptable. Many viewed such harassment, whether face to face or digital, as just a regular part of the school day.

Unsurprisingly, one of the main calls from this report was for compulsory relationships and sex education (RSE) which was contemporary in nature. However, the government response to this call was, once again, that this was not necessary and it was down to schools to manage both RSE and the broader topic of Personal, Social and Health Education (PSHE) as part of the wider school responsibility around maintaining a broad and balance curriculum. It has been argued that such issues are now explored within the OFSTED framework (http://swgfl.org.uk/news/News/online-safety/Making-Sense-of-the-New-Online-Safety-Standards) and therefore senior leaders will respond to the inspection guidance and ensure such education is in place.

However, a recent report by the British Humanist Society (https://humanism.org.uk/wp-content/uploads/2017-01-25-FINAL-Healthy-Happy-Safe.pdf) which analysed over 2,000 recent OFSTED inspections noted that PSHE in general is rarely addressed in these reports, and technologically related social issues are covered particularly poorly, with only 3% of reports mentioning sexting at all.

Conclusion

In conclusion, it would seem the recent comments by Mr Hunt around children's mental health and the influence of digital technology follow a well-trodden path by politicians to lay responsibility at the door of service providers, to make claims that they are responsible for such problems and therefore should "do more" to prevent them. Similar demands have been made of service providers around issues such as children's access to pornography, cyberbullying and radicalisation. However, with issues such as sexting, the focus of Mr Hunt's comments in this part of his evidence to the Select Committee, there is only so much a service provider can do. With concepts like indecency difficult to define for human intervention, we can see an even greater challenge for a technological intervention. Moreover, the flaws in such technologically mediated interventions can result in serious compromises to the rights of the child and to the welfare of families. While the criminal justice response to the challenge of sexting is improving, both in terms of the protection of victims and the pragmatism exercised in the interpretation of relevant law in sexting cases, it still does not provide an effective solution. Equally, the solution cannot be said to reside completely in social responses to the problem, yet the one area where there is potential to achieve great impact is in education. The Government could have some positive and wide ranging influence on the problem of sexting in focusing policy development on a coordinated educational response, yet still it chooses to avoid engaging educational interventions with this problem. We would suggest that rather than finger pointing at other stakeholders in this important area of child safety and wellbeing, it is time the Government "did more" to solve the problem itself.

References

Peersman, C., Schulze, C., Rashid, A., Brennan, M., & Fischer, C. "iCOP: Live forensics to reveal previously unknown criminal media on P2P networks." *Digital Investigation* 18 (2016): 50-64.

Phippen, A. "Children's Online Behaviour and Safety: Policy and Rights Challenges". Palgrave, UK. (2017)

Ringrose, J., Gill, R., Livingstone, S., Harvey, L. "A Qualitative Study of Children, Young People and 'Sexting'" (2012). London: NSPCC.

Rui, Yong, Thomas S. Huang, and Shih-Fu Chang. "Image retrieval: Current techniques, promising directions, and open issues." *Journal of visual communication and image representation* 10.1 (1999): 39-62.