Mass Surveillance and Artificial Intelligence

New Legal Challenges

John Danaher, NUI Galway

In the mid-19th century, a set of laws were created to address the menace that newly-invented automobiles and locomotives posed to other road users. One of the first such laws was the English *The Locomotive Act 1865*, which subsequently became known as the 'Red Flag Act'. Under this act, any user of a self-propelled vehicle had to ensure that at least two people were employed to manage the vehicle and that one of these persons:

> "*while any locomotive is in motion, shall precede such locomotive on foot by not less than sixty yards, and shall carry a red flag constantly displayed, and shall warn the riders and drivers of horses of the approach of such locomotives…"*

The motive behind this law was commendable. Automobiles did pose a new threat to other, more vulnerable, road users. But to modern eyes the law was also, clearly, ridiculous. To suggest that every car should be preceded by a pedestrian waving a red flag would seem to defeat the point of having a car: the whole idea is that it is faster and more efficient than walking. The ridiculous nature of the law eventually became apparent to its creators and all such laws were repealed in the 1890s, approximately 30 years after their introduction.[1]

The story of the Red Flag laws shows that legal systems often get new and emerging technologies badly wrong. By focusing on the obvious or immediate risks, the law can neglect the long-term benefits and costs.

I mention all this by way of warning. As I understand it, it has been over 20 years since the Law Reform Commission considered the legal challenges around privacy and surveillance. A lot has happened in the intervening decades. My goal in this talk is to give some sense of where we are now and what issues may need to be addressed over the coming years. In doing this, I hope not to forget the lesson of the Red Flag laws.

1. What's changed?

Let me start with the obvious question. What has changed, technologically speaking, since the LRC last considered issues around privacy and surveillance? Two things stand out.

First, we have entered an era of *mass surveillance*. The proliferation of digital devices — laptops, computers, tablets, smart phones, smart watches, smart cars, smart fridges, smart thermostats and so forth — combined with increased internet connectivity has resulted in a world in which we are all now monitored and recorded every minute of every day of our lives. The cheapness and ubiquity of data collecting devices means that it is now, in principle, possible to imbue every object, animal and person with some data-monitoring technology. The result is what some scholars refer to as the 'internet of everything' and with it the possibility of a perfect 'digital panopticon'. This era of mass surveillance puts increased pressure on privacy and, at least within the EU, has prompted significant legislative intervention in the form of the GDPR.

Second, we have created technologies that can take advantage of all the data that is being collected. To state the obvious: data alone is not enough. As all lawyers know, it is easy to befuddle the opposition in a complex law suit by 'dumping' a lot of data on them during discovery. They drown in the resultant sea of information. It is what we do with the data that really matters. In this respect, it is the marriage of mass surveillance with new kinds of artificial intelligence that creates the new legal challenges that we must now tackle with some urgency.

Artificial intelligence allows us to do three important things with the vast quantities of data that are now being collected:

(i) It enables new kinds of pattern matching - what I mean here is that AI systems can spot patterns in data that were historically difficult for computer systems to spot (e.g. image or voice recognition), and that may also be difficult, if not impossible, for humans to spot due to their complexity. To put it another way, AI allows us to *understand* data in new ways.

(ii) It enables the creation of new kinds of informational product - what I mean here is that the AI systems don't simply rebroadcast, dispassionate and objective forms of the data we collect. They actively construct and reshape the data into artifacts that can be more or less useful to humans.

(iii) It enables new kinds of action and behaviour - what I mean here is that the informational products created by these AI systems are not simply inert artifacts that we observe with bemused detachment. They are prompts to change and alter human behaviour and decision-making.

On top of all this, these AI systems do these things with increasing autonomy (or, less controversially, automation). Although humans do assist the AI systems in both understanding, constructing and acting on foot of the data being collected, advances in AI and robotics make it increasingly possible for machines to do things without direct human assistance or intervention.

It is these ways of using data, coupled with increasing automation, that I believe give rise to the new legal challenges. It is impossible for me to cover all of these challenges in this talk. So what I will do instead is to discuss three case studies that I think are indicative of the kinds of challenges that need to be addressed, and that correspond to the three things we can now do with the data that we are collecting.

2. Case Study: Facial Recognition Technology

The first case study has to do with facial recognition technology. This is an excellent example of how AI can understand data in new ways. Facial recognition technology is essentially like fingerprinting for the face. From a selection of images, an algorithm can construct a unique mathematical model of your facial features, which can then be used to track and trace your identity across numerous locations.

The potential conveniences of this technology are considerable: faster security clearance at airports; an easy way to record and confirm attendance in schools; an end to complex passwords when accessing and using your digital services; a way for security services to track and identify criminals; a tool for locating missing persons and finding old friends. Little surprise then that many of us have already welcomed the technology into our lives. It is now the default security setting on the current generation of smartphones. It is also being trialled at airports (including Dublin Airport),[2] train stations and public squares around the world. It is cheap and easily plugged into existing CCTV surveillance systems. It can also take advantage of the vast databases of facial images collected by governments and social media engines.

Despite its advantages, facial recognition technology also poses a significant number of risks. It enables and normalises blanket surveillance of individuals across numerous environments. This makes it the perfect tool for oppressive governments and manipulative corporations. Our faces are one of our most unique and important features, central to our sense of who we are and how we relate to each other — think of the Beatles immortal line ‘Eleanor Rigby puts on the face that she keeps in the jar by the door’ — facial recognition technology captures this unique feature and turns into a digital product that can be copied and traded, and used for marketing, intimidation and harassment.

Consider, for example, the unintended consequences of the FindFace app that was released in Russia in 2016. Intended by its creators to be a way of making new friends, the FindFace app matched images on your phone with images in social media databases, thus allowing you to identify people you may have met but whose names you cannot remember. Suppose you met someone at a party, took a picture together with them, but then didn’t get their name. FindFace allows you use the photo to trace their real identity.[3] What a wonderful idea, right? Now you need never miss out on an opportunity for friendship because of oversight or poor memory. Well, as you might imagine, the app also has a dark side. It turns out to be the perfect technology for stalkers, harassers and doxxers (the internet slang for those who want to out people’s real world identities). Anyone who is trying to hide or obscure their identity can now be traced and tracked by anyone who happens to take a photograph of them.

What's more, facial recognition technology is not perfect. It has been shown to be less reliable when dealing with non-white faces, and there are several documented cases in which it matches the wrong faces, thus wrongly assuming someone is a criminal when they are not. For example, many US drivers have had their licences cancelled because an algorithm has found two faces on a licence database to be suspiciously similar and has then wrongly assumed the people in question to be using a false identity. In another famous illustration of the problem, 28 members of the US congress (most of them members of racial minorities), were falsely matched with criminal mugshots using facial recognition technology created by Amazon.[4] As some researchers have put it, the widespread and indiscriminate use of facial recognition means that we are all now part of a perpetual line-up that is both biased and error prone.[5] The conveniences of facial recognition thus come at a price, one that often only becomes apparent when something goes wrong, and is more costly for some social groups than others.

What should be done about this from a legal perspective? The obvious answer is to carefully regulate the technology to manage its risks and opportunities. This is, in a sense, what is already being done under the GDPR. Article 9 of the GDPR stipulates that facial recognition is a kind of biometric data that is subject to special protections. The default position is that it should not be collected, but this is subject to a long list of qualifications and exceptions. It is, for example, permissible to collect it if the data has already been made public, if you get the explicit consent of the person, if it serves some legitimate public interest, if it is medically necessary or necessary for public health reasons, if it is necessary to protect other rights and so on. Clearly the GDPR does restrict facial recognition in some ways. A recent Swedish case fined a school for the indiscriminate use of facial recognition for attendance monitoring.[6] Nevertheless, the long list of exceptions makes the widespread use of facial recognition not just a possibility but a likelihood. This is something the EU is aware of and in light of the Swedish case they have signalled an intention to introduce stricter regulation of facial recognition.

This is something we in Ireland should also be considering. The GDPR allows states to introduce stricter protections against certain kinds of data collection. And, according to some privacy scholars, we need the strictest possible protections to to save us from the depredations of facial recognition. Woodrow Hartzog, one of the foremost privacy scholars in the US, and Evan Selinger, a philosopher specialising in the ethics of technology, have recently argued that facial recognition technology must be banned. As they put it (somewhat alarmingly):[7]

> “ *The future of human flourishing depends upon facial recognition technology being banned before the systems become too entrenched in our lives. Otherwise, people won’t know what it’s like to be in public without being automatically identified, profiled, and potentially exploited.”*

They caution against anyone who thinks that the technology can be procedurally regulated, arguing that governmental and commercial interests will always lobby for expansion of the technology beyond its initially prescribed remit. They also argue that attempts at informed consent will be (and already are) a ‘spectacular failure’ because people don’t understand what they are consenting to when they give away their facial fingerprint.

Some people might find this call for a categorical ban extreme, unnecessary and impractical. Why throw the baby out with the bathwater and other cliches to that effect. But I would like to suggest that there is something worth taking seriously here, particularly since facial recognition technology is just the tip of the iceberg of data collection. People are already experimenting with emotion recognition technology, which uses facial images to predict future behaviour in real time, and there are many other kinds of sensitive data that are being collected, digitised and traded. Genetic data is perhaps the most obvious other example. Given that data is what fuels the fire of AI, it is possible that we should consider cutting off some of the fuel supply in its entirety.

3. Case Study: Deepfakes

Let me move on to my second case study. This one has to do with how AI is used to create new informational products from data. As an illustration of this I will focus on so-called 'deepfake' technology. This is a machine learning technique that allows you to construct realistic synthetic media from databases of images and audio files. The most prevalent use of deepfakes is, perhaps unsurprisingly, in the world of pornography, where the faces of famous actors have been repeatedly grafted onto porn videos. This is disturbing and makes deepfakes an ideal technology for 'synthetic' revenge porn.

Perhaps more socially significant than this, however, are the potential political uses of deepfake technology. In 2017, a team of researchers at the University of Washington created a series of deepfake videos of Barack Obama which I will now play for you.[8] The images in these videos are artificial. They haven't been edited together from different clips. They have been synthetically constructed by an algorithm from a database of audiovisual materials. Obviously, the video isn't entirely convincing. If you look and listen closely you can see that there is something stilted and artificial about it. In addition to this it uses pre-recorded audio clips to sync to the synthetic video. Nevertheless, if you weren't looking too closely, you might be convinced it was real. Furthermore, there are other teams working on using the same basic technique to create synthetic audio too. So, as the technology improves, it could be very difficult for even the most discerning viewers to tell the difference between fiction and reality.

Now there is nothing new about synthetic media. With the support of the New Zealand Law Foundation, Tom Barraclough and Curtis Barnes have published one of the most detailed investigations into the legal policy implications of deepfake technology.[9] In their report, they highlight the fact that an awful lot of existing audiovisual media is synthetic: it is all processed, manipulated and edited to some degree. There is also a long history of creating artistic and satirical synthetic representations of political and public figures. Think, for example, of the caricatures in *Punch* magazine or in the puppet show *Spitting Image*. Many people who use deepfake technology to create synthetic media will, no doubt, claim a legitimate purpose in doing so. They will say they are engaging in legitimate satire or critique, or producing works of artistic significance.

Nevertheless, there does seem to be something worrying about deepfake technology. The highly realistic nature of the audiovisual material being created makes it the ideal vehicle for harassment, manipulation, defamation, forgery and fraud. Furthermore, the realism of the resultant material also poses significant epistemic challenges for society. The philosopher Regina Rini captures this problem well. She argues that deepfake technology poses a threat to our society's 'epistemic backstop'. What she means is that as a society we are highly reliant on testimony from others to get by. We rely on it for news and information, we use it to form expectations about the world and build trust in others. But we know that testimony is not always reliable. Sometimes people will lie to us; sometimes they will forget what really happened. Audiovisual recordings provide an important check on potentially misleading forms of testimony. They encourage honesty and competence. As Rini puts it:[10]

> "*The availability of recordings undergirds the norms of testimonial practice…Our awareness of the possibility of being recorded provides a quasi-independent check on reckless testifying, thereby strengthening the reasonability of relying on the words of others. Recordings do this in two distinctive ways: actively correcting errors in past testimony and passively regulating ongoing testimonial practices."*

The problem with deepfake technology is that it undermines this function. Audiovisual recordings can no longer provide the epistemic backstop that keeps us honest.

What does this mean for the law? I am not overly concerned about the impact of deepfake technology on legal evidence-gathering practices. The legal system, with its insistence on 'chain of custody' and testimonial verification of audiovisual materials, is perhaps better placed than most to deal with the threat of deepfakes (though there will be an increased need for forensic experts to identify deepfake recordings in court proceedings). What I am more concerned about is how deepfake technologies will be weaponised to harm and intimidate others — particularly members of vulnerable populations. The question is whether anything can be done to provide legal redress for these problems? As Barraclough and Barnes point out in their report, it is exceptionally difficult to legislate in this area. How do you define the difference between real and synthetic media (if at all)? How do you balance the free speech rights against the potential harms to others? Do we need specialised laws to do this or are existing laws on defamation and fraud (say) up to the task? Furthermore, given that deepfakes can be created and distributed by unknown actors, who would the potential cause of action be against?

These are difficult questions to answer. The one concrete suggestion I would make is that any existing or proposed legislation on 'revenge porn' should be modified so that it explicitly covers the possibility of synthetic revenge porn. Ireland is currently in the midst of legislating against the nonconsensual sharing of 'intimate images' in the Harassment, Harmful Communications and Related Offences Bill. I note that the current wording of the offence in section 4 of the Bill covers images that have been 'altered' but someone might argue that synthetically constructed images are not, strictly speaking, altered. There may be plans to change this wording to cover this possibility — I know that consultations and amendments to the Bill are ongoing[11] — but if there aren't then I suggest that there should be.

To reiterate, I am using deepfake technology as an illustration of a more general problem. There are many other ways in which the combination data and AI can be used to mess with the distinction between fact and fiction. The algorithmic curation and promotion of fake news, for example, or the use of virtual and augmented reality to manipulate our perception of public and private spaces, both pose significant threats to property rights, privacy rights and political rights. We need to do something to legally manage this brave new (technologically constructed) world.

4. Case Study: Algorithmic Risk Prediction

Let me turn turn now to my final case study. This one has to do with how data can be used to prompt new actions and behaviours in the world. For this case study, I will look to the world of algorithmic risk prediction. This is where we take a collection of datapoints concerning an individual's behaviour and lifestyle and feed it into an algorithm that can make predictions about their likely future behaviour. This is a long-standing practice in insurance, and is now being used in making credit decisions, tax auditing, child protection, and criminal justice (to name but a few examples). I'll focus on its use in criminal justice for illustrative purposes.

Specifically, I will focus on the debate surrounding the COMPAS algorithm, that has been used in a number of US states. The COMPAS algorithm (created by a company called Northpointe, now called Equivant) uses datapoints to generate a recidivism risk score for criminal defendants. The datapoints include things like the person's age at arrest, their prior arrest/conviction record, the number of family members who have been arrested/convicted, their address, their education and job and so on. These are then weighted together using an algorithm to generate a risk score. The exact weighting procedure is unclear, since the COMPAS algorithm is a proprietary technology, but the company that created it has released a considerable amount of information about the datapoints it uses into the public domain.

If you know anything about the COMPAS algorithm you will know that it has been controversial. The controversy stems from two features of how the algorithm works. First, the algorithm is relatively opaque. This is a problem because the fair administration of justice requires that legal decision-making be transparent and open to challenge. A defendant has a right to know how a tribunal or court arrived at its decision and to challenge or question its reasoning. If this information isn't known — either because the algorithm is intrinsically opaque or has been intentionally rendered opaque for reasons of intellectual property — then this principle of fair administration is not being upheld. This was one of the grounds on which the use of COMPAS algorithm was challenged in the US case of *Loomis v Wisconsin.*[12] In that case, the defendant, Loomis, challenged his sentencing decision on the basis that the trial court had relied on the COMPAS risk score in reaching its decision. His challenge was ultimately unsuccessful. The Wisconsin Supreme Court reasoned that the trial court had not relied solely on the COMPAS risk score in reaching its decision. The risk score was just one input into the court's decision-making process, which was itself transparent and open to challenge. That said, the court did agree that courts should be wary when relying on such algorithms and said that warnings should be attached to the scores to highlight their limitations.

The second controversy associated with the COMPAS algorithm has to do with its apparent racial bias. To understand this controversy I need to say a little bit more about how the algorithm works. Very roughly, the COMPAS algorithm is used to sort defendants into to outcome ‘buckets’: a high risk reoffender bucket or a low risk reoffender bucket. A number of years back a group of data journalists based at ProPublica conducted an investigation into which kinds of defendants got sorted into those buckets. They discovered something disturbing. They found that the COMPAS algorithm was more likely to give black defendants a false positive high risk score and more likely to give white defendants a false negative low risk score. The exact figures are given in the table. Put another way, the COMPAS algorithm tended to rate black defendants as being higher risk than they actually were and white defendants as being lower risk than they actually were. This was all despite the fact that the algorithm did not explicitly use race as a criterion in its risk scores.

Needless to say, the makers of the COMPAS algorithm were not happy about this finding. They defended their algorithm, arguing that it was in fact fair and non-discriminatory because it was well calibrated. In other words, they argued that it was equally accurate in scoring defendants, irrespective of their race. If it said a black defendant was high risk, it was right about 60% of the time and if it said that a white defendant was high risk, it was right about 60% of the time. This turns out to be true. The reason why it doesnt immediately look like it is equally accurate upon a first glance at the relevant figures is that there are a lot more black defendants than white defendants -- an unfortunate feature of the US criminal justice system that is not caused by the algorithm but is, rather, a feature the algorithm has to work around.

So what is going on here? Is the algorithm fair or not? Here is where things get interesting. Several groups of mathematicians analysed this case and showed that the main problem here is that the makers of COMPAS and the data journalists were working with different conceptions of fairness and that these conceptions were fundamentally incompatible. This is something that can be formally proved. The clearest articulation of this proof can be found in a paper by Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan.□[13] To simplify their argument, they said that there are two things you might want a fair decision algorithm to do: (i) you might want it to be well-calibrated (i.e. equally accurate in its scoring irrespective of racial group); (ii) you might want it to achieve an equal representation for all groups in the outcome buckets. They then proved that except in two unusual cases, it is impossible to satisfy both criteria. The two unusual cases are when the algorithm is a perfect predictor (i.e. it always get things right) or, a lternatively, when the base rates for the relevant populations are the same (e.g. there are the same number of black defedants as there are white defendants). Since no algorithmic decision procedure is a perfect predictor, and since our world is full of base rate inequalities, this means that no plausible real-world use of a predictive algorithm is likely to be perfectly fair and non-discriminatory. Whats more, this is generally true for all algorithmic risk predictions and not just true for cases involving recidivism risk. If you would like to see a non-mathematical illustration of the problem, I highly recommend checking out a recent article in the MIT Technology Review which includes a game you can play using the COMPAS algorithm and which illustrates the hard tradeoff between different conceptions of fairness.[14]

What does all this mean for the law? Well, when it comes to the issue of transparency and challengeability, it is worth noting that the GDPR, in articles 13-15 and article 22, contains what some people refer to as a ‘right to explanation’. It states that, when automated decision procedures are used, people have a right to access meaningful information about the logic underlying the procedures. What this meaningful information looks like in practice is open to some interpretation, though there is now an increasing amount of guidance from national data protection units about what is expected.[15] But in some ways this misses the deeper point. Even if we make these procedures perfectly transparent and explainable, there remains the question about how we manage the hard tradeoff between different conceptions of fairness and non-discrimination. Our legal conceptions of fairness are multidimensional and require us to balance competing interests. When we rely on human decision-makers to determine what is fair, we accept that there will be some fudging and compromise involved. Right now, we let this fudging take place inside the minds of the human decision-makers, oftentimes without questioning it too much or making it too explicit. The problem with algorithmic risk predictions is that they force us to make this fudging explicit and precise. We can no longer pretend that the decision has successfully balanced all the competing interests and demands. We have to pick and choose. Thus, in some ways, the real challenge with these systems is not that they are opaque and non-transparent but, rather, that when they are transparent they force us to make hard choices.

To some, this is the great advantage of algorithmic risk prediction. A paper by Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan and Cass Sunstein entitled 'Discrimination in the Age of the Algorithm' makes this very case.□[16] They argue that the real problem at the moment is that decision-making is discriminatory and its discriminatory nature is often implicit and hidden from view. The widespread use of transparent algorithms will force it into the open where it can be washed by the great disinfectant of sunlight. But I suspect others will be less sanguine about this new world of algorithmically mediated justice. They will argue that human-led decision-making, with its implicit fudging, is preferable, partly because it allows us to sustain the illusion of justice. Which world do we want to live in? The transparent and explicit world imagined by Kleinberg et al, or the murky and more implicit world of human decision-making? This is also a key legal challenge for the modern age.

5. Conclusion

It's time for me to wrap up. One lingering question you might have is whether any of the challenges outlined above are genuinely new. This is a topic worth debating. In one sense, there is nothing completely new about the challenges I have just discussed. We have been dealing with variations of them for as long as humans have lived in complex, literate societies. Nevertheless, there are some differences with the past. There are differences of *scope* and *scale* — mass surveillance and AI enables collection of data at an unprecedented scale and its use on millions of people at the same time. There are differences of *speed* and *individuation* — AI systems can update their operating parameters in real time and in highly individualised ways. And finally, there are the crucial differences in the degree of *autonomy* with which these systems operate, which can lead to problems in how we assign legal responsibility and liability.

[1] I am indebted to Jacob Turner for drawing my attention to this story. He discusses it

in his book *Robot Rules - Regulating Artificial Intelligence* (Palgrave MacMillan, 2018). This is probably the best currently available book about Ai and law.

[2] See https://www.irishtimes.com/business/technology/airport-facial-scanning-dystopian-nightmare-rebranded-as-travel-perk-1.3986321; and https://www.dublinairport.com/latest-news/2019/05/31/dublin-airport-participates-in-biometrics-trial

[3] https://arstechnica.com/tech-policy/2016/04/facial-recognition-service-becomes-a-weapon-against-russian-porn-actresses/#

[4] This was a stunt conducted by the ACLU. See here for the press release https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

[5] https://www.perpetuallineup.org/

[6] For the story, see here https://www.bbc.com/news/technology-49489154

[7] Their original call for this can be found here: https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66

[8] The video can be found here; https://www.youtube.com/watch?v=UCwbJxW-ZRg; For more information on the research see here: https://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/; https://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf

[9] The full report can be found here: https://static1.squarespace.com/static/5ca2c7abc2ff614d3d0f74b5/t/5ce26307ad4eec00016e423c/1558340402742/Perception+Inception+Report+EMBARGOED+TILL+21+May+2019.pdf

[10] The paper currently exists in a draft form but can be found here: https://philpapers.org/rec/RINDAT

[11] https://www.dccae.gov.ie/en-ie/communications/consultations/Pages/Regulation-of-Harmful-Online-Content-and-the-Implementation-of-the-revised-Audiovisual-Media-Services-Directive.aspx

[12] For a summary of the judgment, see here: https://harvardlawreview.org/2017/03/state-v-loomis/

[13] "Inherent Tradeoffs in the Fair Determination of Risk Scores" - available here https://arxiv.org/abs/1609.05807

[14] The article can be found at this link -

https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

[15] Casey et al 'Rethinking Explainabie Machines' - available here https://scholarship.law.berkeley.edu/btlj/vol34/iss1/4/

[16] An open access version of the paper can be downloaded here https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669