Journal of Philosophical Investigations
Print ISSN: 2251-7960 Online ISSN: 2423-4419
Homepage: https://philosophy.tabrizu.ac.ir

Ethical Problems of Contemporary AI

Donald Gillies

Emeritus Professor at University College London – England. E-mail: donald.gillies@ucl.ac.uk

Article Info

Article type:
Research Article

Article history:
Received 20 November 2025
Received in revised form 23 November 2025
Accepted 29 November 2025
Published online 20 January 2026

Keywords:
AI, Ethical Problems, Deep Learning, Janet Vertesi.

ABSTRACT

Following the advances in AI produced by the deep learning revolution (2012 to present), AI now has a whole range of applications in industry and commerce. Unfortunately, many of these applications are widely considered to be morally wrong. This situation is illustrated by the example of deepfake pornography. Other examples are more nuanced and there is argument about whether they are morally wrong or not. This is illustrated by the example of targeted advertising which has been defended as acceptable. The arguments for and against targeted advertising are discussed. An account is given of an interesting experiment by Janet Vertesi in which she tried to keep her pregnancy a secret from marketing companies and ended up by being suspected of criminal activity. The conclusion is that targeted advertising is bad and should be abolished.

Cite this article: Gillies, D. (2026). Ethical Problems of Contemporary AI. *Journal of Philosophical Investigations*, 19(53), 683-692. https://doi.org/10.22034/jpiut.2026.70341.4344

Publisher: University of Tabriz.

Intruduction

Until the deep learning revolution, which started in 2012, the dominant approach to AI was logic-based AI, where we take logic in a broad sense to include probability. Logic-based AI had many notable successes. It produced the first AI world chess champion, discovered the proof of a mathematical theorem which eminent human mathematicians had tried in vain to prove, discovered some new laws of nature in the natural sciences, and made progress with automating medical diagnosis. Despite these successes, logic-based AI programs did not have many applications in industry and commerce. There were a few such applications. For example, Quinlan (1986, 85) mentions the use of a logic-based AI program (a derivative of ID3) by Westinghouse Electric's Water Reactor Division in a fuel-enrichment application, which boosted the company's revenue by more than ten million dollars per annum. However, these few applications of logic-based AI were relatively unproblematic. The situation changed dramatically with the beginning of the deep learning revolution in 2012. The novel neural networks-based AI turned out to have a whole range of applications to industry and commerce. Unfortunately, many of these applications are widely considered to be morally wrong. Consequently, the philosophy of AI now includes ethics.

I cannot in a short paper give a comprehensive account of all the bad applications of AI as that would fill an entire book, and indeed entire books have already been written on this theme, e.g. Crawford (2021). Instead, in the next section, I will give an illustrative example of a morally bad application of AI.

1. An Illustrative Example. Deepfake Pornography

I have chosen as my illustrative example of a morally wrong application of AI, *deepfake pornography*. This example has a number of features which make it very suitable for our purposes. To begin with, it concerns something which was only made possible by contemporary AI. A deepfake uses generative AI techniques, such as the *generative adversarial networks (gans)*, which were introduced by Ian Goodfellow and his colleagues in 2014, to insert the face of one person onto the body of another person in a video. This enables images to be produced which look exactly like real people and can show these people saying things or performing actions which they never did. Already by 2017, deepfake pornography was circulating on the internet. A typical piece of deepfake pornography might show a famous media star undressing and performing some erotic acts, which of course never occurred in reality. However, deepfake pornography is also widely made showing ordinary people as part of 'revenge porn' or attempts at blackmail.

It is important to note that deepfake pornography is crucially dependent on an advance in AI. Without the development of generative AI, as part of the deep learning revolution, deepfakes could never have been produced. As far as AI ethics is concerned, it seems to us desirable to limit the examples to bad applications of AI, like the present one. This is because there are quite a number of bad applications of mathematics and computing which do not depend on anything to do with AI. This is shown in Cathy O'Neil's admirable 2016 book: *Weapons of Math Destruction*. O'Neil in her work encountered a range of mathematical models of which she says:

> These mathematical models were opaque … And they tended to punish the poor and the oppressed in our society, while making the rich richer.
>
> I came up with a name for these harmful kinds of models: Weapons of Math Destruction, or WMDs for short (O'Neil, 2016, 3).

Now AI programs do involve mathematical models, and so could be examples of WMDs, but there are many WMDs considered by O'Neil in her book which just involve traditional mathematics and are nothing to do with AI. It is better, in order to avoid confusion, to exclude such WMDs from AI ethics.

The next point to be made about deepfake pornography is that there is an almost complete consensus that it is morally bad. In the UK, there have been on television a number of news programmes about deepfake pornography. These have involved interviews with some of the victims, for whom it has been a humiliating and distressing experience, sometimes leading to long lasting mental health problems. Opinions on ethics can differ widely, but it seems impossible to defend such a harmful practice as deepfake pornography, and few to my knowledge attempt to do so.

Yet, which is my third point, the production of deepfake pornography continues unabated despite general agreement that it is wrong and the practice being made illegal in many countries. Much of the mainstream media is now dominated by social media platforms like YouTube and Instagram that do not produce content themselves but rely on independent content creators, who can earn a share of advertising revenue. These platforms do not allow pornography, but there are a number of shadow platforms such as OnlyFans and PornHub that specialise in pornography. It is relatively easy to set up an account and start posting videos to these platforms, including deepfakes. While the terms will normally prohibit deepfakes, the platforms lack the resources of the incentives to screen the huge number of posted videos for deepfakes, thus allowing the practice to go on.

Kant speaks of "the moral law within" and he is not altogether wrong. If some activity is accepted as morally wrong by nearly all members of society, then most people will refrain from that activity. However, there are a number of factors which can silence "the moral law within" in some cases. One of these factors is the possibility of making large profits from the activity generally judged to be morally wrong.

That concludes the analysis of our illustrative example. It was chosen for being a straightforward case. However, other instances of AI applications are more nuanced and there may be disagreement as to whether they are morally bad or not. In the next section, I will consider a case of this sort.

2. Targeted Advertising

For my representative example of an application of AI where there may be some doubt as to whether it is bad or not, I have chosen *targeted advertising*. This choice raises a number of questions which we must first consider. Earlier I insisted that the ethics of AI should deal with genuine AI applications and not with other applications, which, even if they employ mathematics and computing, do not use AI. So, it should first be asked whether targeted advertising does make use of AI. To ensure that this is so, by targeted advertising in this section, I will mean targeted advertising which has been produced using AI. This qualification is not much of a restriction, however. It is indeed possible to do some targeting of advertisements without using AI, but AI has been the key enabling technology which has allowed the enormous expansion of targeted advertising in recent years, and nowadays nearly all targeted advertisements are produced using AI. The procedure goes roughly as follows. Huge databases of consumer behaviour have been collected by recording the online purchases of individuals which are made using their computers or smart phones. In addition, individuals can be tracked via their smart phones and this reveals data about what shops they visit. Deep learning AI systems are trained using this data and come up with models which enable them to predict whether a particular individual is likely to purchase the products of some company and therefore be a suitable target for advertisements for those products. Moreover, further data can establish whether targeted advertisements have been successful by monitoring the percentage of those targeted who have purchased the products advertised. This data can be used to improve the AI models. Naturally this is only a rough sketch of the sophisticated techniques which have been developed in this area, but it leaves no doubt that the use of AI is essential to nearly all targeted advertising. Without AI, the databases recording millions of consumers would be far too large to be processed by hand, while it

is the enormous size of these databases which enables AI systems to learn models which make fairly accurate predictions.

But is targeted advertising really bad? It might be considered by some as ethically quite justifiable. François Chollet in his 2018 book (pp. 11-12) gives a list of 11 breakthroughs achieved by deep learning, "all in historically difficult areas for machine learning". The 9th is :

> Improved ad targeting, as used by Google, Baidu, and Bing (Chollet, 2018, 12).

The fact that Chollet includes this in his list of breakthroughs suggests that he thought this was a good development, or at least that there was nothing particularly wrong with ad targeting.

In fact, in the early days of targeted advertising, it was often argued that it would be a beneficial development, not only for the advertiser, but also for the consumer. This is revealed by Cathy O'Neil in her 2016 book (pp. 68-69). She recalls that, while she was working for the advertising start-up Instant Media, her office was visited by a prominent venture capitalist, who outlined, correctly as it turned out, the brilliant future of targeted advertising. However, this venture capitalist claimed not only that targeted advertising would be profitable, but that it would be welcomed by customers. As O'Neil says:

> He argued that the coming avalanche of personalized advertising would be so useful and timely that customers would welcome it. They would beg for more. As he saw it, most people objected to advertisements because they were irrelevant to them. In the future, they wouldn't be. Presumably, folks in his exclusive demo would welcome pitches tailored to them, perhaps featuring cottages in the Bahamas, jars of hand-pressed virgin olive oil, or time-shares for private jets. And he joked that he would never have to see another ad for the University of Phoenix – a for-profit education factory that appeals largely to the striving (and more easily cheated) underclasses (O'Neil, 2016, 69).

We can illustrate this argument by considering two hypotheticals, but quite realistic, characters – Ms A and Mr B. Ms A is an enthusiast for healthy eating and is particularly fond of Japanese food. Mr B, by contrast, really loves traditional fast-food. For him, the perfect meal might be something like a cheeseburger and French fries, followed by a

chocolatey donut and washed down with a cola drink.[1] Mr B regards Japanese food as being uneatable, while Ms A shudders with horror at what she regards as greasy and disgusting traditional fast-food. She would never allow any of it to pass her lips.
Given this situation, it would obviously be a waste of time to advertise fast-food to Ms A or Japanese restaurants to Mr B. Moreover, it would be very irritating for Ms A to receive advertisements for cheeseburgers and equally irritating for Mr B to receive advertisements featuring the delights of Japanese cuisine. Suitably targeted advertisements would be much better for both Ms A and Mr B. The irritation just mentioned would disappear. Ms A might receive advertisements for Japanese restaurants which she didn't know about but which she would like to try, while Mr B might receive similar advertisements for fast-food outlets.

All this sounds very admirable and harmonious, but of course there might be snags. Mr B might have been warned by his doctor that he is already overweight and risks developing diabetes, heart problems and other unpleasant conditions. The doctor has strongly advised him to give up fast-food and eat healthier, though not necessarily Japanese, food. Mr B is worried about his health and is trying to follow the doctor's advice, but he still suffers from carvings for his beloved fast-food. Naturally the relevant corporations know very well about this craving and will continue to target him with advertisements for fast-food, tempting him to disobey his doctor's advice and risk the onset of some unpleasant disease. There are of course many other similar examples. Someone might have been a compulsive gambler and lost a lot of money. He is now trying to give up gambling, but his past is known to the advertisers, and he may well continue to receive targeted advertisements for online gambling, tempting him to worsen his already very bad situation. Gambling is a very profitable industry. Even Ms A who is purchasing healthy and beneficial products may get irritated by endless advertisements from local supermarkets proclaiming the excellence of their sushi.

These methods are not limited to the advertisements themselves. Most targeted advertising happens on social media platforms such as Facebook, Instagram or Twitter. Their revenue model relies on their audience viewing (and clicking on) as many adverts as possible. This is served by targeting the adverts so that people are more likely to click on the adverts they see, but it is also served by having people stay on the platform for longer so that they see more adverts. This leads the platforms to also target content to

[1] These are all items on the menu of a fast-food outlet near where the author lives in London, UK. The chocolatey donut is shown in the online menu as being covered with chocolate and having chocolate in the hole of the donut.

make it more and more appealing to each individual. Many think that this has led to the rise of smartphone addiction, with more and more appealing content keeping people on the platform. The targeted content that works best is often the most strongly emotional content, leading to the algorithms pushing politically polarising content or to "beauty" content that is leading to insecurity and mental health problems in teenage girls.

These are significant objections to targeted advertising, but there is another extremely serious objection. Targeted advertising is only possible if large quantities of data about an individual's consumer habits are available to commercial firms. Indeed, such data is regularly appropriated by some firms and sold on to others. It has become part of normal commercial transactions. However, the necessary data can only be obtained by surveillance, and many object to such surveillance as intrusive and invasive of their privacy.

It is frequently pointed out that contemporary society is beginning to exhibit many features of the dystopian future society which George Orwell portrayed in his novel 1984. Orwell's novel is in many ways a very inaccurate prediction of the future. It is set in Britain in 1984 and portrays a country which is under a Stalinist regime run by a dictator 'Big Brother' who is supported by the 'thought police'. As it turned out, Britain in 1984 had shifted to the right and was ruled by Margaret Thatcher an advocate of *laissez-faire* capitalism. In this sense, therefore, Orwell's predictions could not have been more inaccurate. However, in his novel the population is subject to mass surveillance using devices which are described as telescreens. These are fixed to the walls of houses, apartments, offices and public spaces, and cannot be turned off except by high-ranking members of the 'inner party'. Telescreens transmit television programmes, but they also enable the thought police to watch what is going on in front of them. Thus, the citizens of Orwell's dystopian society can be observed by the thought police at any time, and indeed never know whether the thought police are watching what they are doing. This mass surveillance is one of the most horrifying features of Orwell's dystopia. Yet it has been largely implemented in contemporary societies, though by a different technological device. We are all surveyed not by telescreens but by smartphones.

There is another point in which Orwell's predictions are not entirely accurate. Orwell imagined the surveillance carried out by the government in order to detect and suppress any dissidents who opposed the regime. No doubt surveillance by smartphone is carried out by some governments on some individuals, but what Orwell did not imagine at all is that surveillance would be caried out by large capitalist corporations with the aim, not of suppressing dissidents, but of selling more of their products. The extent of this

surveillance is really remarkable as is illustrated by an experiment carried out by Janet Vertesi and described in her 2014 article.

Janet Vertesi is a Professor of Sociology at Princeton, specializing in the sociology of science, knowledge and technology. When she became pregnant, she knew that she was liable to become the target of marketing companies. As she says (2014) Prospective mothers are busy making big purchases and new choices (which diapers? Which bottles?) that will become their patterns for the next several years. In this big data era of targeted advertising, detection algorithms sniff out potentially pregnant clients based on their shopping and browsing patterns. It's a lucrative business; according to a report in the *Financial Times*, identifying a single pregnant woman is worth as much as knowing the age, sex and location of up to 200 people.

Janet Vertesi was also, as a sociologist of technology, launching a study of how people keep their personal information on the Internet. This suggested to her the experiment of seeing whether she could go the entire nine months of her pregnancy without letting the marketing companies know she was pregnant. It turned out that this was very difficult and attempting to carry out the experiment actually led to her being suspected of criminal activity!

It was immediately obvious to Janet Vertesi that she couldn't say anything about her pregnancy on social media such as Facebook or Twitter. However, she had also to warn her relatives not to mention it. As she says (2014) But social interactions online are not just about what you say but also what others says about you. One tagged photo with a visible bump and the cascade of "Congratulations!" would let the cat out of the bag. So, when we phoned our friends and families to tell them the good news, we told them about our experiment, requesting that they not put anything about the pregnancy online.

Unfortunately, this request was ignored or forgotten by an uncle who, at seven months, sent Janet Vertesi a Facebook message congratulating her on her pregnancy. She was forced to reply very rudely online denying the truth. Another family member made a similar mistake in a Facebook chat a few weeks later and rather naively said: "I didn't know that a private message wasn't private."

However, Janet Vertesi had to take further steps not involving social media. As she says (2014) Many websites and companies especially baby-related ones, follow you around the Internet. So, I downloaded Tor, a private browser that routes your traffic through foreign servers. While it has a reputation for facilitating illicit activities, I used it to visit BabyCenter.com and to look up possible names. And when it came to shopping, I did all my purchasing – from prenatal vitamins to baby gear and maternity wear – in

cash. No matter how good the deal, I turned down loyalty-card swipes. I even set up an Amazon.com account tied to an email address hosted on a personal server, delivering to a locker, and paid with gift cards purchase with cash.

However, these measures, such as the use of Tor, even though they were designed simply to evade marketing detection, did begin to make it look as if Janet Vertesi was engaged in criminal activity of some kind. This became obvious when she needed to make an important purchase. As she says (2014) But, as I discovered when I tried to buy a stroller, opting out is not only antisocial, but it can appear criminal. For months I had joked to my family that I was probably on a watch list for my excessive use of Tor and cash withdrawals. But then my husband headed to our local corner store to buy enough gift cards to afford a stroller listed on Amazon. There, a warning sign behind the cashier informed him that the store "reserves the right to limit the daily amount of prepaid card purchases and has an obligation to report excessive transactions to the authorities.

Janet Vertesi (2014) concludes No one should have to act as a criminal just to have some privacy from marketers and tech giants. … It's time for a frank public discussion about how to make personal-information privacy … a basic human right, both online and off.

A survey conducted in the United States in 2012, revealed that most Americans agree with Janet Vertesi in seeing targeted advertising as an invasion of privacy. 68% of those surveyed said they are "not okay" with targeted advertising because they do not like having their online behaviour tracked and analysed.[1] This makes perfect sense. How many would be agreeable to personal details about themselves such as the fact they are pregnant, or that they were once a compulsive gambler, or that they have an addiction to traditional fast food to be appropriated by firms for commercial purposes? However, the use of AI to produce targeted advertisements necessarily involves the appropriation of personal details. Our conclusion therefore is that the use of AI to produce targeted advertisements is a bad application of AI. As nearly all targeted advertisements are produced using AI, this amounts to saying that targeted advertisements are bad and should be abolished.

References

Chollet, F. (2018). *Deep Learning with Python*, Manning.

Crawford, K. (2021). *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press.

[1] Source: Wikipedia article on Targeted Advertising.

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, Penguin Edition, 2017.

Orwell, G. (1949). *1984*, Polygon, 2021.

Quinlan, J. (1986). Induction of Decision Trees, *Machine Learning*, 1, pp. 81-106.

Vertesi, J. (2014). My experiment opting out of big data made me look like a criminal, *Time*. May 1. (available online)