Douwe Korff
Emeritus Professor of International Law, London Metropolitan University
Associate, Oxford Martin School, University of Oxford

Tricking an AI system & the Cybercrime Convention

Introduction

In a 2018 paper that was only now brought to my attention, Calo et al. discuss whether "tricking a robot" constitutes hacking.[1] They focus on five scenarios that they analyse under the U.S. Computer Fraud and Abuse Act (CFAA). I thought it would be worthwhile to look at those same scenarios from the perspective of the Council of Europe Cybercrime Convention (also known as the Budapest Convention).[2] (I refer to AI systems rather than robots, but we mean the same.)

The five scenarios

Calo et al. discuss the following five scenarios:

1. Planting adversarial sound commands in ads.

A perpetrator of intimate partner violence buys a local television advertisement in the jurisdiction he suspects his ex now resides. Embedded in the ad is an adversarial sound input that no person would recognize as meaningful. The attack causes his ex's personal assistant in range of the TV to publish her location on social media.

2. Causing a car crash by defacing a stop sign to appear like a speed limit.

An engineer extensively tests the detector used by the driverless cars company where she works. She reports to the founder that she's found a way to knowingly [and I assume physically – DK] deface a stop sign to trick the car into into accelerating instead of stopping. The founder suspends operations of his own fleet but defaces stop signs near his competitor's driverless car plant. A person is injured when a competitor driverless car misses a stop sign and collides with another vehicle.

3. Shoplifting with anti-surveillance makeup.

An individual steals from a grocery store equipped with facial recognition cameras. In order to reduce the likelihood of detection, the individual wears makeup she understands will make her look like another person entirely to the machine learning model. However, she looks like herself to other shoppers and to grocery store staff.

4. Poisoning a crowd-sourced credit rating system.

A financial start up decides to train an ML model to detect "risky" and "risk averse" behavior so as to assign creditworthiness scores. A component of the model invites internet users to supply and rate sample behaviors on a scale from risky to risk averse. A group of teenagers poison the

---

[1] Calo, Ryan and Evtimov, Ivan and Fernandes, Earlence and Kohno, Tadayoshi and O'Hair, David, Is Tricking a Robot Hacking? (March 27, 2018). University of Washington School of Law Research Paper No. 2018-05, Available at SSRN: https://ssrn.com/abstract=3150530 or http://dx.doi.org/10.2139/ssrn.3150530

[2] Council of Europe Convention on Cybercrime, Budapest, 23 November 2001, CETS 185, available at: https://rm.coe.int/1680081561
There are two protocols to this Convention: the first concerns the criminalisation of acts of a racist and xenophobic nature committed through computer systems (ETS No. 189), and the second facilitates enhanced co-operation and disclosure of electronic evidence (CETS No. 224). But these are not relevant here.

model by supplying thousands of images of skateboarders and rating them all as risk averse. One teenager from the group whose social network page is full of skateboarding pictures secures a loan from the start up and later defaults.

5. Data inversion across international borders.

A European pharmaceutical company trains and releases a companion model with a drug it produces that helps doctors choose the appropriate dosage for patients. The model is trained on European data but subsequently released to doctors in the United States. A malicious employee in the U.S. with access to the model uses an algorithm to systematically reconstruct the training set, including personal information.

Crimes under the Cybercrime Convention

The Cybercrime Convention requires state parties to that convention to create the following main offences in their domestic law:

To intentionally and without right:

- **access** the whole or any part of a computer system (Article 2);
- **intercept**, by technical means, a non-public transmissions of computer data to, from or within a computer system (Article 3);
- damage, delete, deteriorate, alter or suppress computer data (Article 4), referred to in the heading to that article as **data interference**;
- seriously hinder the functioning of a computer system by inputting, transmitting, damaging, deleting, deteriorating, altering or suppressing computer data (Article 5), referred to in the heading to that article as **system interference**.

With regard to the first two offences (access and interception), the Convention stipulates that "[a] Party may require that the offence be committed with dishonest intent, or in relation to a computer system that is connected to another computer system"; and with regard to the third offence (data interference), that state parties may limit that offence to situations in which "serious harm" is caused.

Article 6 sets out a number of ancillary offences that I shall not discuss here.

Article 1 defines the concepts of "computer system" and "computer data" as follows:

> "**computer system**" means any device or a group of interconnected or related devices, one or more of which, pursuant to a program, performs automatic processing of data.
>
> "**computer data**" means any representation of facts, information or concepts in a form suitable for processing in a computer system, including a program suitable to cause a computer system to perform a function.

Assessment

In my opinion, the first attack scenario discussed by Calo et al., planting a command in an advertisement to make a personal assistant disclose computer data, arguably can be said to involve "accessing" a computer system (i.e., that personal assistant), but in any case involves

"inputting" computer data (i.e., the command) into a computer system (i.e., again, that personal assistant), and in the scenario this is done both intentionally and without right (the attacker does not have a right to make their ex's personal assistant/computer system do things without the consent of his ex partner).

The second scenario, defacing a stop sign, is tricky. At first glance, it would appear that the physical defacing of the stop signs does not involve accessing, intercepting or interfering with computer data (but see below).

Does it constitute "system interference"? The defacing of the signs clearly results in the malfunctioning of a computer system, i.e., of the competitor's car computer system. But is this malfunctioning caused by "inputting, transmitting, damaging, deleting, deteriorating, altering or suppressing computer data"?

In my opinion, odd though it may seem, it can be said that – at least in this scenario – the stop signs constitute "computer data" within the meaning of the Convention because they contain and send out "information … in a form suitable for processing in a computer system, including a program suitable to cause a computer system to perform a function". To a computer system that can read traffic signs, a traffic sign constitutes "computer data".

If I am right on this point, the physical defacing of the stop signs constitutes "accessing" and "interfering with" "computer data" (i.e., with the sign that a computer [*in casu* the car computer] can process) and also "system interference" because the defaced sign sends false information to the competitor's car computer. And this is done intentionally and without right.

Can this line of arguing be carried over to the third scenario, in which a shoplifter wears makeup that fools a facial recognition system? Does the thief "deface" her face to send erroneous "computer data" (i.e., her face) to the facial recognition system? In the example, the thief does so intentionally, to "trick" the system. (Another person, who is also not recognised because of her makeup, but who had no such specific intention, would clearly not fall within the ambit of the offence, for which intent is a necessary element.) But is the thief "without [a] right" to make herself unrecognisable to a facial recognition system? We do not yet live in a society where we are legally obliged to make ourselves recognisable to such systems at all times. We may need to be authenticated by biometric systems in some contexts,[3] but that is far from a general duty – if anything, European data protection authorities stress the exceptionality of biometric recognition systems and the European Parliament is proposing a ban on the use of face recognition systems in public places that would include, e.g., shopping centres. Shops may ban entry to people wearing motorbike helmets, and presumably, if they could detect makeup that makes someone unrecognisable to a facial recognition system, they might be able to do the same in that respect. But that does not detract from the fact that we, the public, do have a right to wear motorbike helmets and balaklavas, or face recognition-defying makeup. So I conclude that this scenario does not relate to something that must be criminalised under the Cybercrime Convention.

In the fourth scenario, the model poisoners do not "access", "intercept" or "damage, delete, deteriorate, alter or suppress" pre-existing "computer data". However, in my opinion, they do

[3] The European Data Protection Board has given access to a nuclear facility as an example.

"input" data into the system (the model), and they do so intentionally in order to make the system not function as it should. Nor do I believe that they do this "with right": yes, the system developer (rather stupidly) allows them to enter data into their model, but he does not give them the right to deliberately mislead the model. In other words, in my opinion, the model poisoners' actions constitute "system interference".

Finally, there is the fifth scenario: the unauthorised re-identification of supposedly de-identified data. In this case, the malicious employee did have a right of access to the data – but not a right to re-identify the individuals. In my opinion, the re-identification involved "altering" the data: a data entry that says "The patient with patient number 1234567 was diagnosed with diabetes on 10/05/20" is different from "The patient with patient number 1234567=Mr John Blogs was diagnosed with diabetes on 10/05/20". The malicious employee also made the system perform an action – the re-identification – that was not intended: he made the system malfunction. And he did this both intentionally and without right. In fact, they – and indeed the company they work for – are in clear breach of European data protection law.

In sum: In my opinion, the activities described in scenarios 1, 2, 4 and 5 must in principle be criminalised in all state parties to the Cybercrime Convention, subject only to the permitted qualifications that states may introduce under Article 6 in relation to dishonest intent, connected computer systems, and "serious harm". But the actions of the makeup wearer in the third scenario need not – and in my opinion should not – be made a criminal offence under the Convention.

Of course, in order to get the full picture, one would have to study the laws of the parties to the Cybercrime Convention, which may well not be fully in line with what I conclude above, but I will leave that to others. Calo and his colleagues have already done a great job in relation to US law in that regard.[4]

- o – O – o -

Douwe Korff (Prof.)
Cambridge (UK), 25 September 2023

[4] The USA is a party to the Convention, and has used the option in Article 4 to declare that it "reserves the right to require that the conduct result in serious harm, which shall be determined in accordance with applicable United States federal law." See:
https://www.coe.int/en/web/conventions/cets-number-/-abridged-title-known?module=signatures-by-treaty&treatynum=185 (list of state parties)
https://www.coe.int/en/web/conventions/cets-number-/-abridged-title-known?module=declarations-by-treaty&numSte=185&codeNature=2&codePays=USA (US reservations)