Situational Awareness and Adversarial Machine Learning – Robots, Manners, and Stress

Ross Anderson
University of Cambridge
and University of Edinburgh
ross.anderson@cl.cam.ac.uk

Ilia Shumailov
Computer Laboratory
University of Cambridge
ilia.shumailov@cl.cam.ac.uk

***Abstract*—Both humans and animals infer intent – a dog knows the difference between a kick and a stumble. Over thousands of generations, we have evolved biological and cultural mechanisms to quickly assess the threat posed by another human or an animal, and animals who interact with humans have similar mechanisms. We also have a keen awareness of whether our environment is friendly or hostile. As robots, and other automata that rely on machine learning, become widespread, they will raise similar but more complex questions of signaling and detecting intent. In recent research, we have been exploring how adversarial samples can be detected more easily than they can be blocked, allowing systems to fall back to more cautious modes of operation. The interaction between machine-learning components and service-denial attacks is a fascinating subject that few have studied so far. In short, while classical system resilience may be seen in terms of layered defence and redundancy, that of machine-learning systems may be much more human. Combining the two intelligently could be a new frontier for research, with a focus on situational awareness. We may see new security protocols, as the communication of intent may become more important than the communication of identity not only for the security and safety of interaction between humans and robots, but also between robots and the wider environment. This gives a new and perhaps more realistic angle on both robot ethics and adversarial machine learning.**

***Index Terms*—Autonomous vehicles, robots, drones, affect recognition, geofencing**

While walking down the street near the UC Berkeley campus, one of the authors (Ilia) found himself sharing the pavement with a food-delivery drone (see fig. 1) [12]. As he walked up to the robot and tried to pass it, it suddenly accelerated and started bumping into him. He tried to avoid it but regardless of what he tried, the robot kept up its attack. Feeling overwhelmed, he retreated from the conflict by stepping off the pavement.

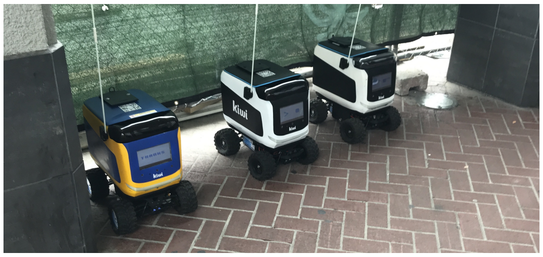

Fig. 1. Kiwi Robot in UC Berkeley

The interaction felt like an encounter with an aggressive stray dog. Unlike dogs, though, drones do not come with non-verbal cues that we humans have learned to read through millions of years of evolution. The world of information security was no help either; there was no authentication protocol to notify the drone that Ilia felt threatened – and fighting down the urge to defend himself. In fact, there was a face screen and it was smiling while attacking! Should we be wary of delivery robots?

Airborne delivery drones have been used for some years in specialised applications, ranging from military supplies in Afghanistan to blood products in Rwanda. They came to broader public attention in 2013 when Jeff Bezos claimed that we were four or five years from drone-at-home delivery. Although Prime Air is not yet available, patents have been filed on how to build drone hives and capture data by scanning people's homes [9]. Amazon claimed that they have only two real obstacles left: technological (e.g. battery life) and legislative. In 2019, new aviation regulations came into force in the UK, removing the latter.

When talking of the drone project, Bezos boasted about safety: 'This thing can't land on somebody's head while they're walking around their neighborhood.' He also mentioned that when a drone can't land or feels nervous, a human operator will be patched in. This isn't entirely reassuring, as the failure of expected human intervention has led to at least one fatal accident involving self-driving cars, but it does acknowledge that situational awareness matters for drones too.

So what does it actually mean to be safe around robots, drones and ML systems in general? Is it only about avoiding fatal accidents? We argue that the problem is much larger. Can we trust robots? Can they trust us? And can they trust each other?

I. Manners for robots

There is a large space between courteous behaviour and deadly assault, for both humans and animals. Dogs bark and growl; personal space may be invaded or respected; intent can be signaled in numerous ways large and small. And so it shall surely be with robots. They can encroach on people's property or personal space in ways ranging from mild to infuriating, and people's responses can range from avoidance to gunfire – via everything in between. One common means of defence or protest against a street delivery drone is to tip it over on its side, leaving it helpless [13]; that this happens, suggests that Ilia's experience was not unique.

The law on trespass gives another starting point. There can be trespass to property (if I enter your land without your consent or a good reason); to goods (if I mess with your stuff); and to the person – the 'infringement of a person's personal integrity'. Trespass to the person includes battery (the wrong I commit if I hit you), but is wider; silence that causes fear, such as silent phone calls, can count; so can aggressive words, and unlawful restraint, or an accidental contact that is not rectified.

A third starting point is deception, which interacts with both manners and the law in complex ways. We tell little white lies all the time to lubricate social exchanges, and to mitigate or excuse a minor trespass ("Oh, did I stand on your foot? I'm dreadfully sorry!"). A world of robots will surely be deceptive at many levels: it will be natural for Amazon to build robots that are small and cute so people will like them rather than attack them. Assassin robots will be small and cute too. And this is nothing new; James Bond is attractive enough (both physically and in terms of personal charm and fake ID) to get inside the perimeter. His robot successors will surely try to be as good.

However most threats are dirtier and low-grade, rather than lethal assassins aimed at major state targets. It's the low-grade stuff that we deal with using the everyday cultural machinery of manners, norms and reputations – aided by the fact that human intent leaks into human behaviour except where people are taking care to deceive. Intent is not straightforward with robots, though, as the intent is in some sense that of the programmer. A company whose products cause personal harm or property damage can be sued or prosecuted. But intent and deception can interact in new ways.

Robots may try to pass as humans. In May 2018 Google rolled out its automated call assistant, Duplex, which can make automated restaurant bookings and arrange appointments. Such conversation bots – let's call them ConBots – may find many useful applications, but may also enable criminals to innovate and scale up attacks ranging from unwanted commercial messaging to finance scams and spear-phishing. The details may depend on whether the victim is duped into believing they are talking to a person, or simply bullied into doing what an apparently authoritative phone call demands. (We've all been trained to realise that if we refuse to tell our mother's maiden name to the lady from the bank we'll end up with dead cards and be unable to buy our groceries.)

One pushback against interlocutors of uncertain humanity is to invoke manners. Indeed, phrases such as 'I think that's a bit impertinent of you' or 'Do you MIND?' may function as rough-and-ready CATPCHAs.

Our starting point, though, is robots that are clearly identifiable as such, whether they are delivery drones, flying surveillance cameras or autonomous vehicles. And even if the typical robots are cute smiling sociopaths owned by profit-maximising tech firms, it will be in the firms' interest to mask the sociopathy and provide mechanisms whereby the robots that interact with human society can negotiate their way in it. A firm operating delivery drones that trundle along the pavement – and occasionally bump into pedestrians – had better work out how to make the drones say sorry; if it does not, it will risk having them kicked over and left helpless, or even banned from the pavements.

The need for robot manners is a potential show-stopper for self-driving cars. While advanced driver assistance systems can operate vehicles on the freeway, towns are much harder. There are many traffic situations where drivers give way to one another in order to keep the traffic flowing, and robocars find these difficult [11]. It's been reported that the Waymo approach can lead to cars taking a long time to execute turns across traffic, inflicting frustration and anger on human drivers of following vehicles. Merging is hard, as you need to know the other vehicles' intentions, and to signal your own intentions to others so that they can open a gap to let you in. Signals include not just explicit ones, such as indicators or flashing headlights, but more subtle ones from small changes of speed to eye contact. And the other parties are not just vehicles; at an intersection, a robocar may have to accommodate cyclists, pedestrians and animals too. Complex city intersections may be one make-or-break case for autonomous vehicles.

Another is congested residential streets where parked cars force drivers to give way to oncoming traffic. These are not just difficult for robot drivers; they can tax human drivers who are inexperienced, impatient or elderly. And while the manners of a pavement delivery drone can perhaps involve patching in a human operator in the event of a collision, this won't work for suburban driving. The reaction time of human safety drivers is inadequate, even if they're in the vehicle – let alone if they're in a call centre hundreds of miles away.

II. Robots safe for mixed environments

So what is meant by security and safety in the context of robots that interact with the public?

The classic concepts of security engineering and safety engineering can give us some of our requirements. We don't want a drone to be exploited remotely by a bad actor who causes it to kill people, whether at random in a terrorist attack, or in a more targeted way. We don't want a mechanical failure to cause harm either; airborne drones might have parachutes. We don't want widely-deployed civilian robots to be so vulnerable to well-known military technologies (such as jamming and GPS spoofing) that in time of tension an adversary could block a nation's streets with immobilized robotaxis[1]. Our threat model for autonomous vehicles, and for robots in general, will include all of the above.

However, this will not be enough. We have to think of interaction with the public. Until now, security engineers tended to think only of a 'user', and we don't always deal with users effectively, as researchers from Whitten and Tygar onwards have taught us [22]. Sometimes we assume that some of the users are hostile (as with ATMs) and sometimes that there are wicked insiders (as in the design of HSMs). However

[1] Some rental cars will not start if they can't get a mobile phone signal

we have usually avoided thinking about 'the public' because of the starting assumption that the first thing we do is to authenticate people.

Now consider the street traffic in Naples, or in Delhi. People simply walk out into the street and raise their hand to slow down oncoming cars (in the event that the traffic is moving quickly enough to present a hazard). A robotaxi in such a city cannot authenticate every driver with whom it interacts, let alone every elephant, dog or cow[2]. The realistic decision facing the robo-driver is whether to stop, or to proceed slowly with the horn blaring.

Like it or not, robots will have to accommodate the local human culture. It's not just vehicle behaviour, but pedestrian behaviour too; the norms for acceptable interpersonal distance are not the same in the USA, Israel and India. There may be ethical issues for international firms as local norms of precedence and dominance may be dependent on physical size, gender, age, dress and much else. The engineer will also have to be sensitive to such moral factors as care, harm, authority and sanctity.

We will also have to think about scalable citizen attacks on robots, as there have already been a number of incidents in California and Arizona where angry citizens have obstructed self-driving cars, slashed their tyres or pelted them with rocks [13]. Meanwhile security robots have been covered in tarpaulins or had their sensors covered with sauce [23]. What else can we anticipate? Might there be some adversarial image which, if printed on a T-shirt, will cause drones to avoid me, or even crash? Or would people chalk signs on the road to confuse them, or set up cardboard policemen? Such tactics might appeal to student pranksters, but serious protesters – such as delivery drivers afraid for their jobs – might just walk in front of robot vehicles and cause them to stop.

It is clearly preferable if people who object to robot behaviour can use due process. But what might the mechanics look like? It might be reasonable to expect (or even require) robots to display their operator's brand name, just as many commercial delivery vehicles do. So if I see Starship robot 117 behaving badly, I can broadcast this fact. But how do I know its number? Must it display ”117” on its shoulder, like a vehicle license plate? In the case of drones, governments are moving: the USA regulated for the remote identification of unmanned aircraft in January 2021. New drones must broadcast their position using 'drone remote ID' within 18 months of that date, and existing devices must be retrofitted within 30 months.

In any case, I'd rather not protest via an ecosystem controlled by the robot's owner or vendor, but through an external channel such as Twitter. Citizen videos have been a boon in holding bad cops to account. So do ground-based robots need licence plates backed up with criminal penalties, as cars already have? Do we need behavioural traps for other robots, just like we have car speed cameras? Do we need drones to advertise their presence, such as by requiring a robot vehicle to have a flashing green light?

III. How might interaction with the public develop?

Prudent drivers usually steer clear of cars that behave aggressively, although there are always a few who get into road-rage incidents. Robot drivers – and delivery robots – currently give way to aggressive humans and this is likely to continue, at least outside of law-enforcement and military applications. So automatic human affect recognition starts to matter. At the laboratory scale, we have some working prototypes: a whole range of different biological markers, from facial cues to body movements, can be used to infer human mental states [3], [14], [15], [17]. Direct affect recognition is feasible at short range; some cars already detect if their drivers are tired, and they could also detect aggression and try to calm down the driver. However, affect detection at a distance from human faces and body language remains problematic, not just technically but ethically too [4]. This has driven engineers to study the behavioural classification of vehicles as opposed to drivers. Google already filed a patent on a 'Method to Detect Nearby Aggressive Drivers and Adjust Driving Modes' in 2013 [5], and indeed the first granted US patent for the automatic detection of aggression in nearby vehicles was filed in the last century, in 1999 [10].

This might perhaps extend in the future to how delivery robots can detect aggressive pedestrians. But given the difficulty of remote affect recognition, it might be simpler to have a hand sign convention for a pedestrian to tell a drone to go away. Civilian airborne drones must already have geofencing to stop them wandering into prohibited or restricted airspace[3]; why not a law that drones should back off from individuals who signal a trespass of their property, their privacy or their personal space?

And even if we solve the problem of robot-to-human understanding, it will not solve the more complex problem of human understanding, or misunderstanding, of robot intent. It is well known that humans tend to associate human-like feelings to non-human objects (anthropomorphism) and second, that humans tend to see patterns where they do not exist (pareidolia). Anthropomorphism and pareidolia, coupled with the importance of non-verbal cues in human communication [8], leave us a hard problem with no obvious answer.

The typical citizen now spends more time being told what to do by computers – as we fill out endless forms and wrestle with badly-designed automation. In future, will we be physically pushed about by robots too?

IV. Adversarial machine learning

The above discussion of safety shows the importance to both humans and robots of being able to discern each others' intent, both to reassure others and to detect hostile actors. This

[2] In those Indian states with a cow-murder law, killing a cow can put the vehicle occupants in peril of their lives, as well as the engineer if he were available for extradition.

[3] That geofencing can be defeated with about the same amount of effort as is needed to root a smartphone is neither here nor there; a drone that intrudes on Gatwick airport is clear evidence of criminal intent.

brings us to security, which is increasingly intertwined with safety in the world of robots and of the 'Internet of Things' in general. Where robots, or other devices or systems, rely on machine learning to understand their environment, we rapidly come to the problem of adversarial machine learning. Just as robots cheat humans by pretending to be cute and friendly, so also can humans cheat robots, using ever more sophisticated techniques. And it is to be expected that robots (and other ML systems) will be designed to cheat each other.

Machine learning started being used in spam filters in the 1990s, and the spammers promptly invented a range of tricks to game them [7]. Many of these tricks have gone across to, or been reinvented since, the neural network revolution [2]. You can poison the training data, such as by using captive accounts to mark spam as 'not spam'; in general, if a model continues to train itself in use, it can sometimes be simple to lead it astray. As with any learning, the syllabus determines the outcome. You can attack the inference phase and cause the model to give the wrong answer by perturbing the input in a way that maximises the prediction error; people found they could use gradient-descent methods to find imperceptible input changes that would cause images to be wildly misclassified [21]. You can also do service-denial attacks by using similar methods to select inputs that cause the model to take the maximum amount of time, or energy, to classify them. While straight-through image classification tasks can be slowed down perhaps 20%, the more complex pipelines used in natural language processing are extremely vulnerable to such attacks, and can be slowed down by factors of hundreds or even thousands of times. Automated techniques exist to find such attacks efficiently; they often come up with odd things, such as inserting a few Chinese characters in a Russian text, which could in theory be spotted and used to sound an alarm [18].

A very general lesson learned from this first round of attack and defence co-evolution is that it's much easier to detect attacks and respond using other mechanisms, than to try to build ML models that are of themselves robust to adversarial inputs. In the case of service-denial attacks, for example, one can either design the model for worst-case performance, or set a hard limit on the amount of time and energy it will expend on trying to understand any input. If we consider an NLP system in isolation, then its failure on certain inputs might seem like a bug. But if it's a component of a larger system that can respond in other ways to attack, fragility can become a feature. It will be the canary in the mine – the early-warning mechanism to put a system on alert.

Another example of useful fragility comes from adversarial attacks on machine-vision systems. At present, the systems used in advanced driver assistance systems in cars are detuned so as to make them less vulnerable to adversarial images. Other approaches include training systems with a large number of adversarial samples to make them more robust; this can help manage sensitivity to small changes, but it imposes real performance constraints, and will still be vulnerable to different adversarial examples.

An alternative we have developed is the Taboo Trap [19]. The underlying insight is that when you first send your child to school, it is well-spoken with beautiful manners; yet within a week or two it's starting to use words of which your grandmother might not approve. In short, exposure to adversarial samples can be detected via a breach in the social taboos with which the child was trained.

This idea can be adapted to detect adversarial samples in ML systems. We select some set of outputs, or of intermediate activation values, and declare them to be taboo. This set forms a 'key' that customises an instance of a model during training, in the same way that a cryptographic key can customise a communication. We then train our model to avoid these taboo activations and outputs. If the model later produces one, this signals that it has encountered inputs that were not in its training set; and past a certain threshold, these signals can be interpreted as an alarm. As different instances of the model can be trained against different keys, an attacker who develops adversarial instances against one instance cannot rely on their working against other instances [16], [20].

Machine-learning techniques have spread beyond spam filters to anti-virus scanners, intrusion detection systems and many other security tools. A real problem for the security industry is that professional malware developers now test their products against the standard defences, and if these are static and deterministic it may be fairly easy to tweak the malware until it can get through. Some diversity of defence may help prevent attacks from scaling.

It is difficult to make a stand-alone machine-learning system robust against adversarial inputs. ML models absolutely have to deal with out-of-distribution data. That makes robustness a rather misaligned effort – it is better to flag up as anomalous a datapoint that is not a cat or a dog (such as a picture of a dog with a cat tattoo). But that brings the problem of defining non-anomalous behaviour – the general problem that intrusion detection systems try to solve in the first place!

A further point is that very few components are stateful in ML, which is both nice and terrible. State is hard to handle, and once you do it can easily make things even more brittle – so that random noise breaks your model as well as sophisticated attacks [24]. So robotic arms that plan their trajectories and adapt will experience both safety and security problems in adversarial settings. Feedback is sometimes essential, but it increases complexity (it breaks composability in even the simplest of security models) [1].

So what is the way forward? The critical change of perspective is to think in terms of designing systems using multiple components, some of them with machine learning. The designer must then work out what to do when an attack is detected on one of its machine-learning models. In the case of a car, this may mean sounding an alarm and coming safely to a stop. Indeed, some countries already require an autonomous vehicle to have a safety kernel that, in the event of confusion, will cause it to slow down and stop without changing lane. Even this is a lot harder than it looks, as vehicles often get confused when going through road works and don't know which lane they're in!

We humans have evolved mechanisms that make us wary when we meet a rival, and alert when in a context that we feel to be dangerous. Stress hormones increase heart rate and respiration, while muscles tense; this ‘fight-or-flight’ response evolved so we can react quickly to life-threatening situations. But alertness is expensive, and when triggered inappropriately (say by work pressure or family difficulties), chronic stress can lead to high blood pressure, anxiety, depression, and addiction. We are particularly vulnerable to stressors that suggest hostile intent, that violate tribal taboos, where the level of the risk is uncertain, or that leave us feeling we’re not in control [1].

Conversely, when we feel safe and reassured, we relax. An interesting aspect of this is the placebo effect. When we truly believe that we’re safer and expect to get better, we can relax and put more effort into recovering from disease or injury, thereby hastening the expected result. These effects are genuine, powerful and widespread, providing the physiological basis for faith healing. They are now understood to be an evolved system for bodily resource management; once your subconscious is convinced it doesn’t need to husband resources in case something worse turns up, it will throw much more energy into a full-scale immune response [6].

It should surprise no-one that as digital systems start to use neural-network and other machine-learning mechanisms inspired by human cognition, they will increasingly need to be aware whether they are under attack, or indeed whether an attack may be imminent. Adversarial samples are a signal that a robot or other ML system is under attack.

However dealing with real-life risks and threats could well be more complex for robots than it was in our ancestral evolutionary environment. Our early ancestors could climb a tree when they saw a lion; but if an adversarial image causes cars to slow down and stop, it could cause chaos with urban traffic.

As with the African savannah 200,000 years ago, deception is likely to be pervasive. Just as the great majority of animals are trying to eat and to avoid being eaten, and have evolved a multitude of deceptive strategies to help them, so also the great majority of robots (and other ML systems) will be operated by profit-maximising companies. Some will be operated by governments, but may be no less sociopathic for that.

However, there are three significant differences. First, we have rapid broadcast communications: the first time an Amazon drone drops a package on a toddler’s head, we will all know about it. Second, although the large companies and governments that operate most robots and other ML systems may be sociopathic in their disregard for individuals, they operate within the rule of law and must at least pretend to care. Third, the modern world is data-rich, with billions of people carrying smartphones, and with other specialist data collection devices such as dashcams, which many professional drivers operate for insurance purposes.

In time we will no doubt evolve laws and social norms appropriate for a world with pervasive machine-learning systems. Meanwhile, the prospect of crowdsourced accountability based on pervasive logging, holding out the threat of brand damage, and backed up in the most egregious cases by the existing civil and criminal law, will have to restrain the worst excesses.

V. Conclusion

Manners matter. Much interpersonal communication is non-verbal, and indeed some evolutionary psychologists believe that the origins of human speech lie in rhythmic chanting developed to align emotions for hunting or conflict. Yet current research in robotics sort-of assumes that while affective computing is a nice sensor to have, it is unproblematic in security, safety and ethical terms. A philosopher might even ask whether it is meaningful for robots to have manners? Or is being a robot so unlike being a human – even more unlike than being a bat – that it just makes no sense?

Yet robot manners are starting to have real functional importance. Whether robot cars and human-driven ones can share complex intersections is an important test case; whether humans will happily share pavements with delivery drones and even security drones is another.

The manners we display when moving around are a kind of security protocol, but one radically different from those studied by cryptographers. The identification and authentication of individual human and robot actors is almost irrelevant (except for post-facto forensics and justice). The signaling of aggression is a tacit negotiation that animals undertake when moving around and staying out of each others’ way. The signaling of intent, in general, is of huge importance, and has been ignored for too long. When a lot of the players are robots, whose intent is in some sense deceptive, functional communication may depend on at least the possibility of surveillance.

Tacit signalling, whether of aggression or of another intent, is in turn part of a much larger problem, namely situational awareness. Robots and other systems using machine learning are vulnerable to a range of attacks, notably adversarial samples, which are rare yet powerful. The costs of immunising ML systems against such attacks is in many cases unreasonably high. The most practical defence will often be an alarm: to detect the presence of such samples and fall back to more conservative modes of operation. This is similar to animals’ fight-or-flight reaction. And just as the inappropriate triggering of this reaction by the stresses of modern life can lead to anxiety, depression and other stress-related illness, so also ML systems might conceivably be stressed if inappropriate inputs become pervasive.

This leads to a new possible attack, which we might call *stress jamming* – where an opponent floods and environment with adversarial samples in order to force robots and other ML systems into degraded modes of operation. Until now, researchers have thought of adversarial attacks as targeted; but as many such samples are transferible [16], they might also be used at scale. We leave this thought as a challenge for future research.

Acknowledgements

The second author was supported with funds from Bosch-Forschungsstiftung im Stifterverband. Particular thanks to Richard Clayton, Darija Halatova, Benjamin Maximilian Reinheimer and Domagoj Stolfa for useful discussions.

References

[1] R. J. Anderson. *Security engineering: A guide to building dependable distributed systems*. John Wiley & Sons, 2020.

[2] B. Biggio and F. Rolli. Wild patterns: Ten years after the rise of adversarial machine learning. 2018.

[3] O. Celiktutan and H. Gunes. Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. *IEEE Transactions on Affective Computing*, 8(1):29–42, Jan 2017.

[4] K. Crawford. Time to regulate ai that interprets human emotions. *Nature*, 592(167), 2021.

[5] P.-Y. Droz and J. Zhu. US Patent Application 13/772,615. 21 Feb 2013.

[6] N. Humphrey. Great expectations: The evolutionary psychology of faith-healing and the placebo effect. In *The Mind Made Flesh: Essays from the Frontiers of Evolution and Psychology*, 2002.

[7] A. Hutchings, S. Pastrana, and R. Clayton. Displacing big data. In *The Human Factor of Cybercrime*, 2020.

[8] M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[9] S. Levin. Amazon patents beehive-like structure to house delivery drones in cities, 2018.

[10] J. J. Mackey and D. M. Mackey. Us Patent 6392564b1. 10 May 1999.

[11] A. Marshall. How robo-cars handle the frustratingly human act of merging. *Wired*, 12 Dec 2018.

[12] C. McDonald. Forget the robot apocalypse. order lunch., 2018.

[13] T. Mogg. Waymo's autonomous cars are coming under attack in arizona. *Digital Trends*, 1 Jan 2019.

[14] J. Ondras and H. Gunes. Detecting deception and suspicion in dyadic game interactions. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 200–209, New York, NY, USA, 2018. ACM.

[15] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, 5:705–721, 2017.

[16] I. Shumailov, X. Gao, Y. Zhao, R. Mullins, R. Anderson, and C.-Z. Xu. Sitatapatra: Blocking the transfer of adversarial samples. 2019.

[17] I. Shumailov and H. Gunes. Computational analysis of valence and arousal in virtual reality gaming using lower arm electromyograms. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pages 164–169. IEEE, 2017.

[18] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson. Sponge examples: Energy-latency attacks on neural networks. In *6th IEEE European Symposium on Security and Privacy (EuroS&P)*, 2021.

[19] I. Shumailov, Y. Zhao, R. Mullins, and R. Anderson. The taboo trap: Behavioural detection of adversarial samples. *arXiv preprint arXiv:1811.07375*, 2018.

[20] I. Shumailov, Y. Zhao, R. Mullins, and R. Anderson. Towards certifiable adversarial sample detection. *arXiv preprint arXiv:2002.08740*, 2020.

[21] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

[22] A. Whitten and D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. *8th Usenix Security Symposium*, 1999.

[23] J. C. Wong. Rage against the machine: self-driving cars attacked by angry Californians. *The Guardian*, 6 Mar 2018.

[24] Y. Zhao, I. Shumailov, H. Cui, X. Gao, R. Mullins, and R. Anderson. Blackbox attacks on reinforcement learning agents using approximated temporal information. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pages 16–24, 2020.