Please notice that this is a DRAFT of a paper the final version of which has been published as Fossa, F., Paparusso, L., Braghin, F. (2024). “Interaction Attacks as Deceitful Connected and Automated Vehicle Behavior”, in Parkinson, S., Nikita, A., Vallati, M. (Eds), *Deception in Autonomous Transport Systems. Threats, Impacts and Mitigation Policies*, Springer, Cham, pp. 147-162. https://doi.org/10.1007/978-3-031-55044-7_10

Interaction Attacks as Deceitful Connected and Automated Vehicle Behaviour

Fabio Fossa[1], Luca Paparusso[2] and Francesco Braghin[3]

[1], Department of Mechanical Engineering, Politecnico di Milano, via La Masa 1, 20156 Milano, Italy
fabio.fossa@polimi.it
[2], Department of Mechanical Engineering, Politecnico di Milano, via La Masa 1, 20156 Milano, Italy
luca.paparusso@polimi.it
[3], Department of Mechanical Engineering, Politecnico di Milano, via La Masa 1, 20156 Milano, Italy
francesco.braghin@polimi.it

Abstract. The present chapter aims at exploring the idea of *interaction attacks* as a form of *deceitful* Connected and Automated Vehicle (CAV) behaviour that requires to be adequately counteracted both on the technical and social level. After some introductory remarks on cyberattacks, deception, and driving automation, we argue that interaction attacks and related risks still require to be adequately conceptualised. To this aim, we draw on Norbert Wiener’s notes on animals and cybernetic systems to show that the possibility of interaction attacks based on deceptive behaviour stems from the very nature of control in machines. Using Wiener’s insights and recent literature as a blueprint, we then provide a conceptual description of interaction attacks involving CAVs. In addition, we discuss a case study aimed at further clarifying the phenomenon. Finally, we advance some remarks on interaction attacks as a form of deceitful CAV behaviour according to the framework elaborated by [3] and call for further research on such a critical issue.

Keywords: Driving Automation, Interaction Attacks, Deception.

1 Introduction

Connected and Automated Vehicles (CAVs) are expected to massively revolutionise the transportation world. While purported benefits have been soon identified and often exaggerated, potential risks are increasingly capturing researchers' attention. Among others, risks linked to malicious attacks exploiting various vulnerabilities have raised several cybersecurity concerns. Well-founded fears of attacks putting road users in danger have shed a suspicious light on the promises of driving automation in terms of safety and reliability, thus clarifying that ethically relevant benefits will come not as simple by-products of CAV development and deployment, but rather only as actively pursued social objectives [1].

Building on this presupposition, the present chapter explores the idea of interaction attacks [2] as a form of deceitful CAV behaviour that requires to be adequately counteracted both on the technical and social level. Section 2 offers some introductory remarks on cyberattacks, deception, and driving automation to argue that interaction attacks and related risks still require to be adequately conceptualised. To this aim, section 3 draws on Norbert Wiener's notes on animals and cybernetic systems to show that the possibility of interaction attacks based on deceptive behaviour stems from the very nature of control in machines. Using Wiener's insights and recent literature as a blueprint, Section 4 provides a conceptual description of interaction attacks involving CAVs. Section 5 discusses a case study aimed at further clarifying the phenomenon and demonstrating its practical feasibility. Finally, Section 6 concludes the chapter by advancing some preliminary remarks on interaction attacks as a form of deceitful CAV behaviour according to the framework elaborated by [3] and calling for further research on such a critical issue.

2 Driving Automation, Cyberattacks, and Deception

When complex artificial systems are to be deployed in close vicinity to human beings, safety issues are necessarily to be assessed with due care. Among the related preoccupations, the possibility that malicious agents could intentionally interfere with the expected functioning of the systems must be considered. To minimize the risk of external attackers taking control of system operations, adequate cybersecurity countermeasures must be taken.

CAVs raise several cybersecurity challenges. As an innovative vehicle technology combining mechanical engineering and computer science, driving automation must cope with safety issues present in both fields. In addition to more traditional considerations concerning reliability and crashworthiness, vulnerabilities proper to digital systems must be adequately addressed.

As a matter of fact, cybersecurity represents a major concern in the field driving automation. Worries that driving automation might be exposed to attacks putting the safety of passengers and road users in jeopardy are widely acknowledged. Just as any other online computing device, CAVs are susceptible to various cyberattacks [4]. The complexity of driving automation systems – which could be perhaps better defined as systems of systems – is bound to present numerous vulnerabilities. Their malicious exploitation could pose tremendous safety threats. Through cyberattacks, vehicles could be remotely tampered with or taken control of and hijacked to cause massive traffic disruption or harm passengers and other road users [5, 6, 7, 8, 9]. Similar events would pose significant threats to social safety. Concurrently, they would likely undermine social trust in the technology and lead to widespread rejection [10, 11].

Being so closely connected to the protection of road users' physical integrity, cybersecurity is of evident ethical interest. Safety is a well-established value in the realm of transportation. Stakeholders have a clear obligation to reduce safety risks by any reasonable means. Actually, the ethical significance of ensuring high levels of cybersecurity for CAVs is so widely acknowledged that little need has been felt in the literature of justifying its moral relevance. The identification, prevention, and mitigation of cybersecurity risks are commonly recognized as crucial obligations that engineers, designers, programmers, manufacturers, etc. have a duty to satisfy. Key safety-enhancing cybersecurity values such as robustness, resilience, and integrity have been adopted in the context of driving automation as well [12, 13]. Accordingly, various cybersecurity risks proper to CAVs are being systematically identified and strategies are being introduced to prevent them or to mitigate negative outcomes [14, 15].

In line with the focus on the digital element of cybersecurity, most of the identified threats involve attacks to various system components through software manipulation or sensor interference. For instance, Rizvi and colleagues [16] have shown that Denial-of-Service attacks can allow for taking remote control of brakes, acceleration, and steering. Moreover, sensors like radar, LiDAR, and cameras are variously vulnerable to the emission of signals intentionally aimed at impairing their functionality or deceiving them into perceiving non-existent objects [17]. Furthermore, systems can also be attacked more indirectly by intentionally tampering with elements of the environment. For instance, ultrasonic sensors are vulnerable to cloaking attacks, where obstacles covered with sound-absorbing materials are made undetectable [18]. Similarly, traffic signs can be modified to confuse machine vision algorithms and, possibly, influence CAV behaviour in predetermined ways [19].

According to Nikitas and colleagues [3], many categories of cyberattacks– spanning from spoofing and flooding to Denial-of-Service and Man-in-the-Middle attacks – exhibit an element of *deception*. However, it is important to be clear on the extent to which the notion of deception applies here. Consider, e.g., attacks targeting sensing technologies. In a sense, tampering with sensors intends to deceive systems into believing that certain objects are there, while they are not; or that certain objects are not there, while they actually are. This sounds very similar to how the deception of a human driver could be described. However, can driving systems be deceived just as human drivers can?

Arguably, it would be too anthropomorphic to state that driving systems can be deceitfully induced to form and act on false beliefs, as human drivers could. Such account, which involves mental states and beliefs, is unnecessarily complex and human-like. A simpler conceptualisation would suffice to provide a clear understanding of our case. When driving systems are said to be deceived, what is meant is that a divergence is observable between system data on a given state of affairs on the one hand, and the actual state of affairs on the other. A divergence, moreover, that cannot be traced back to sensor or algorithmic malfunctioning, but can only be explained by reference to a malicious attack by an antagonistic agent. A machine could then be said to have been deceived when an external agent intentionally and successfully causes a divergence between the data on which the system bases its operations and the actual state of affairs.

In all the cases considered so far, CAVs are victims of deceptive acts external to the domain of driving. Deception comes from agents that do not partake in the driving game. Rather, adversaries launch their attacks while standing off the road. Moreover, deception is not achieved through driving behaviours. On the contrary, systems are deceived through the use of non-CAV technologies directly tampering with their functionalities, interfering with their operations, and disrupting their behaviour.

However, CAVs need not be only targets of external deception. As suggested in [3], CAVs could be "deceitful" too: they could also perform attacks while rolling on the road. More specifically, they could do so by executing *deceptive driving manoeuvres*. They could carry out their attacks not as external agents, then, but as players in the driving game.

In what follows, we focus the attention on the possibility of deceitful CAVs manipulating the behaviour of other CAVs through deceptive *interactions*. The possibility of manipulating the behaviour of artificial systems through artfully designed interaction patterns stems from the very nature of how control and communication in the machine work. Indeed, this eventuality was already foreshadowed by Norbert Wiener, the founding father of cybernetics. To get a clearer idea of the tie that binds interactions, control, and deception, let us now turn to Wiener's insights.

3 Control, feedback, and fakes

In his 1961 essay "On Learning and Self-Reproducing Machines", Norbert Wiener offered extremely insightful suggestions on a particular form of the entanglement between artificial systems and deception. In a sense, argues Wiener, deception is part of the behaviour of cybernetic systems just as it is part of the behaviour exhibited by the entities they simulate – i.e., organisms, animals in particular. Since animals engage in deceptive behaviours, machines could also be made to behave in analogous ways.

Wiener's main purpose in the first part of the essay is to explore the idea of "the learning of game-playing machines which enables them to improve the strategy and tactics of their performance by experience" [20: 170]. If chess-playing machines are considered, Wiener explains, "the mere obedience to the laws of the game" is not what poses the most critical challenges. Rather, it is the computation of the strategy or policy to be followed while playing by the rules that raises the greatest problems. If a given strategy is selected and computed, an expert adversary would soon figure it out, exploit its weaknesses, and consistently win. Variability and adaptation to circumstances is necessary to design a well-functioning artificial chess player. Both could be obtained through learning from previous games. The consideration of past moves and their effectiveness would be fed back on the weights associated to the pieces and their relative positions on the chessboard. This way, the machine would not exhibit the same behaviour over and over again, but would adapt to the strategy followed by its opponent. The logic of learning machines, then, is that of feedback. And since previous games must be analysed and reliable patterns found out of vast amount of data, statistics become a necessary ingredient of programming.

This conceptualisation of learning in game-playing machines, Wiener notes, proves particular enlightening if applied to "the activity of struggle" – i.e., when two or more agents compete against each other in a shared space. In order to support his observation, Wiener considers four examples taken from the lifeworld: the fight between the mongoose and the rattlesnake as experienced by Wiener himself and described by Rudyard Kipling in his short story "Rikki Tikki Tavi" [21]; a roadrunner attacking a rattlesnake as depicted in an old Walt Disney movie; the "dance with death" [20: 174] opposing the bull and the bullfighter; and, finally, various contests where human beings face one another – smallswords duelling, fencing, and a game of tennis. What all these agonic situations share is the interactive logic they are determined by. Each player's strategy aims at putting the opponent in a position of disadvantage and frustrating her effort to control the game according to her own purposes. Each player adapts her strategy on the basis of previous knowledge, previous experience, skills, and her opponent's behaviour.

It is precisely in the space of the *agon* that opportunities for deception arise. Consider the struggle between the mongoose and the rattlesnake. Wiener describes the strategy of the mongoose as a series of *feints*, or *fakes*, aimed at inducing the snake to attack and, thus, occupy a position that is advantageous (in the long run) for the mongoose. In this sense, through its feints the mongoose deceives the rattlesnake into playing a game of moves and countermoves that, in the end, will result in its defeat. By feigning, the mongoose gains an advantage over the rattlesnake which allows for striking the final, lethal attack. As Wiener concludes, "this time the mongoose's attack is not a feint but a deadly accurate bite through the cobra brain". The veil of deception falls and the deceived pays the price of its guilelessness.

Bullfighting, fencing, and tennis can also be described in an analogous way. As long as players engaged in "interlaced coordinated actions", the possibility of deceiving the opponent through feints is there to be exploited. As a matter of fact, in many sports (such as, e.g., basketball and football) feints and fakes constitute a fundamental move type each player must learn to master. And feints are difficult to conceive without a reference to deception. Indeed, feints allow players to exert indirect control over their opponents by deceiving them into thinking that a given course of action is about to happen, while this is not the case. Feints manipulate the predictive performances of opponents, leading them to act in ways that only apparently serve their own purposes. Actually, and inadvertently, they end up serving the faker's purposes. Ensnared by the deceiver, the deceived loses any possibility to counteract efficaciously and becomes vulnerable.

Whether the same conceptual framework of the agonic experience, with its deceptive component, can be applied meaningfully to situations opposing animals against animals, animals against humans, and humans against humans, is likely to be controversial. Deception, make believe, acting as if, etc. all require fine mental capacities that many might deem mistaken to attribute to animals. In line with the behaviourist presupposition of cybernetics [22], however, Wiener suggests not only that this is the case, but that machines too can exhibit this form of behaviour. Indeed, while "the snake's pattern of action is confined to single darts, each one of itself" [20: 174], the mongoose plays a more complex and articulated game based on previous moves, future predictions, and fakes. "To this extent", concludes Wiener, "the mongoose acts like a learning machine" [20: 174] – and humans, behaviourists would add, do the same as well, even if at higher degrees of complexity. *To an extent*, then, fakes and deceptive interactions also belong to the domain of cybernetic systems. Machines can be programmed to (learn to) deceive – i.e., to perform feints and, thus, gain an advantage over their opponent (be it an animal, a human, or another machine), indirectly controlling their behaviour according to given plans. These machines, we surmise, might also be CAVs.

Before discussing the possibility of deceptive interactions between CAVs, a note of caution on language use is necessary. Using the same words to frame machine and organic behaviour can be tricky. It might easily prompt biomorphic attribution of features proper to organic life on to entities that do not belong to the same category. Misattribution of organic-like or human-like qualities on to artificial systems is particularly dangerous in the case of advanced AI systems to which morally significant tasks are delegated. As argued extensively in [1], driving is a moral activity, deeply entangled with crucial ethical values such as responsibility, autonomy, safety, and so on. Misattributing autonomy, responsibility, or even the human capacity of deception to artificial systems is bound to lead to confusion and misunderstandings.

In light of this, in the next section we focus the attention on driving automation and offer further clarification of *the extent to which* the notions of "feint", "fake", and "deception" apply to the interlaced coordinated CAV actions. By doing so, we will be able to precisely outline the concept of interaction attacks and account for the related element of deception.

4 Interaction Attacks

Wiener's observations on system control, feedback, and feints offer an insightful viewpoint from which to explore the possibility of CAVs engaging in deceitful driving behaviour. Indeed, the interactions between CAVs on the road can be construed as a sort of agon, or competition, where different players pursue their own goals while influencing and being influenced by the other vehicles with which they share the road.

The vast majority of the goals pursued are likely to be transport-related: getting to desired destinations. However, this need not always be the case. The goal of a CAV could also be much more worrisome – i.e., attacking other vehicles, causing them to crash or drive off the road. Interactions between CAVs could also become a "dance with death". And deception could turn into a powerful weapon in the hands of attackers. In this section we present the idea of "interaction attack" in driving automation [2] and clarify the extent to which deception could be said to play a part in it.

Interaction attacks are based on the claim that once knowledge is obtained concerning how a system works, it can be gamed – meaning that the logic controlling its behaviour can be intelligently bent according to a plan. This way, control over system behaviour could be exerted indirectly, by exposing the system to circumstances that will elicit the desired reactions. Interaction attacks would not need to violate system security or tamper with sensors for them to malfunction or give false positives and negatives. On the contrary, system behaviour would be manipulated precisely because the logic behind it is known and can be intentionally exploited. These attacks, then, assume that the system keeps working as it is supposed to. They count on it. By involving a CAV into a purposefully designed interaction pattern, attackers can thus influence its behaviour according to their own agenda – just as the mongoose influences the behaviour of the rattlesnake, forces it into a vulnerable position, and strikes.

For these reasons, we believe that interaction attacks could pose a significant threat to driving automation safety. While – at least to our knowledge – the problem is yet to be fully identified and acknowledged in cybersecurity terms, similar problems addressing the interaction between road users and CAVs have been pointed out in the literature on the ethics of driving automation [7, 9, 23, 24, 25]. More specifically, scholars have noted that if CAV safety features were known, pedestrians and other road users might exploit them to get an unfair advantage with reference to the right of way. Moreover, relying on the expected efficiency of these systems, road users might engage in behaviours that would otherwise be considered dangerous. For example, pedestrians could start stepping abruptly and unattentively on the street to cross it, knowing that CAVs will detect them and break. Also, cyclists or motorcyclists could start occupying crossroads regardless of the right of way, knowing that CAVs will put safety first and yield.

The problem is analysed in detail by Millard-Ball [26]. In this article, street crossing is framed as a "game of chicken" where right of way is ultimately determined by an equilibrium between the payoffs that each involved party gets as a result of a given choice. When deciding whether to cross the street, pedestrians consider both the prospected benefits of doing so and the risks they expose themselves to. Risks, of course, are due to the fact that drivers might fail to stop and hit them. Even if it is in the drivers' best interests to yield, they could be distracted, intoxicated, tired, or aggressively asserting right of way. Drivers also carry out similar evaluations to decide whether or not to allow pedestrians to pass. Such a decision-making process must account for many individual and environmental aspects. For instance, considerations concerning the behaviour of the other party must be combined with expectations based on implicit norms and explicit traffic rules. Different areas might in fact be regulated differently. For example, pedestrians and drivers might exercise different degrees of attention or adopt different behaviours in a busy city centre or in a small country village. As a result of this bargaining, either pedestrians or drivers get to pass first.

Driving automation would have a dramatic effect on such equilibrium. Since the CAV control logics would implement risk-averse safety features, the risk of being hit for pedestrians would substantially plunge even when jaywalking. Passengers would not even be able to assert right of way, since the driving system would automatically yield to protect pedestrians' safety. As a result, CAVs would lose much of their attractiveness as means of urban transportation, where interaction with pedestrians and other human road users could hardly be excluded. Therefore, Millard-Ball suggests, policy and regulative frameworks must take this eventuality into consideration and introduce countermeasures to disincentivise over-assertive behaviour on the part of human road users.

Even though the abovementioned cases offer a good basis to conceptualise our problem, they tackle a slightly different issue. First of all, even though the situations considered oppose traffic participants, they portray human taking advantage of CAVs – and not CAVs attacking other CAVs. Moreover, what is at stake is the emergence of behaviours that put in danger those who practice them and that could potentially disrupt automated traffic. System features are maliciously exploited to gain an advantage in terms of road use. Aggressive road users, however, do not have any interest in putting CAV passengers in danger. If anything, they expose themselves to the risk of suffering harm should the system be unable to handle the situation safely. System behaviour, then, is manipulated only momentarily and with no further goals in mind. Nonetheless, this situation could represent the first step towards more articulated manipulation plans, where a series of known system reactions are intelligently stimulated to force the CAV to behave as desired. This is what interaction attacks intend to achieve.

Indeed, and building on Wiener's notes, CAVs could be *designed to deceive*: to fake or feign traffic manoeuvres that would appear to be neither uncommon nor safety-critical at first, but would end up putting a targeted vehicle in danger. Knowledge of the control logics determining the driving behaviour of the attacked vehicle would make it possible to predict its future position as a result of the reactions the attacker forces it to execute. Move after move, manoeuvre after manoeuvre, the attacked vehicle can be lured into dangerous traffic situations without impairing or interfering with its functionalities, but just by maliciously bending its control logics to a hidden agenda.

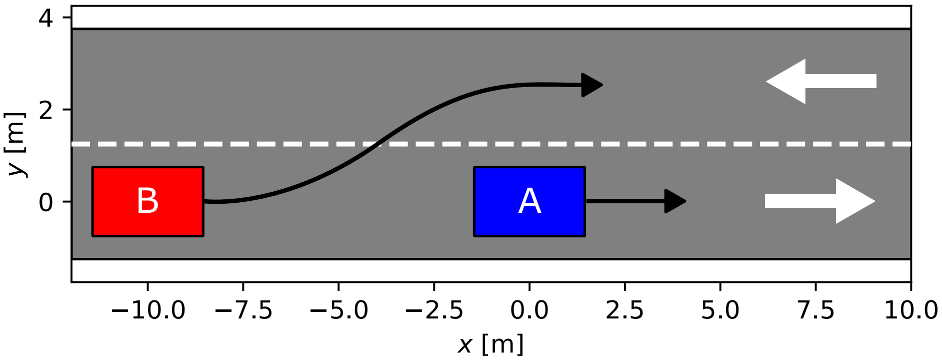
5 A Case Study

In order to further clarify what is meant by interaction attacks and to provide some support to their potential execution, let us introduce the following case study.[1].

[1] The code of this case study can be found online at https://github.com/lpaparusso/overtake_cyberattack

Consider the scenario schematised in Fig. 1. Two CAVs, A and B, drive on a straight two-way road with two lanes, one per each traffic direction. Vehicle A drives one the right lane. Vehicle B follows Vehicle A on the same lane, but at higher speed. Consequently, Vehicle B starts overtaking Vehicle A.

Fig. 1. Schematisation of the case study



At this point, Vehicle A initiates an interaction attack. The objective of Vehicle A is to prevent Vehicle B from having enough free space laterally to return on the right lane, either in front or behind Vehicle A. Due to potential upcoming vehicles in the opposite direction on the left lane, the interaction attack carried out through Vehicle A exposes Vehicle B to a very dangerous situation.

During interaction attacks, control over the behaviour of attacked CAVs is exerted indirectly. This kind of indirect control can be achieved by intentionally designing the behaviour of Vehicle A so to execute manoeuvres that force Vehicle B to occupy the desired position on the road. By forecasting the intent of Vehicle B – or, in other words, artfully manipulating interactions and reactions among the two CAVs – the attack can be successfully carried out.

The problem is now formalised in mathematical notation. The two CAVs are both modelled with a discrete-time single-track kinematic model

$$\begin{cases} x_\mathrm{i}(k+1) = x_\mathrm{i}(k) + v_\mathrm{i}(k)\cos\big(\psi_\mathrm{i}(k)\big)\ \Delta\mathrm{t} \\ y_\mathrm{i}(k+1) = y_\mathrm{i}(k) + v_\mathrm{i}(k)\sin\big(\psi(k)\big)\ \Delta\mathrm{t} \\ \psi_\mathrm{i}(k+1) = \psi_\mathrm{i}(k) + v_\mathrm{i}(k)/\mathrm{L}\tan\big(\mathrm{\delta_i}(k)\big)\ \Delta\mathrm{t} \\ v_\mathrm{i}(k+1) = v_\mathrm{i}(k) + a_\mathrm{i}(k)\,\Delta\mathrm{t} \end{cases} \tag{1}$$

The subscript i can take values in {A, B}, and is used to differentiate the variables corresponding to vehicles A and B. The discrete timestep is denoted with k and the time discretisation is Δt. The state variables are the vehicle coordinates in the global reference frame x and y, the global heading angle ψ, and the speed of the vehicle v. The parameter L defines the distance between the rear and front axle of the vehicle. Finally, the control variables are the steering angle δ and the acceleration a.

The state and control variables are subject to constraints, which express the physical limitations of the vehicle dynamics and actuators. Constraints are formalised as follows:

$$\begin{gathered}
-\delta_{\text{max, i}} \leq \delta_{\text{i}}(k) \leq \delta_{\text{max, i}} \\
v_{\text{min, i}} \leq v_{\text{i}}(k) \leq v_{\text{max, i}} \\
a_{\text{min, i}} \leq a_{\text{i}}(k) \leq a_{\text{max, i}} \\
j_{\text{min, i}} \leq \big(a_{\text{i}}(k+1) - a_{\text{i}}(k)\big)/\,\Delta t \leq j_{\text{max, i}}
\end{gathered} \tag{2}$$

where j represents the jerk of the vehicle, and the subscripts min and max denote the corresponding minimum and maximum allowed values, respectively. In this section, we only detail the behaviour of the two CAVs when the overtaking has already started – i.e., when Vehicle B has just finished its shift towards the left lane, ready to perform the overtaking. Indeed, this is the very moment in which the attack begins.

The steering actions δ_{A} and δ_{B} are provided by a lateral controller for lane keeping. The acceleration a_{B} of Vehicle B, instead, follows a pre-defined behaviour, which is now illustrated. At the beginning of the attack, Vehicle B imposes a_{B} to be

$$a_{\text{B}}(k) = \max_{r \in [a_{\text{min},B},\, a_{\text{max},B}]} r \qquad s.t.\,(2) \tag{3}$$

in order to overtake Vehicle A as fast as possible. Since it would be dangerous to stay too long on the left lane due to possible incoming traffic in the opposite direction, the driving system is designed to shift back towards the right lane after a period of, say, 4 seconds. If overtaking is not completed within 4 seconds the acceleration is then switched to

$$a_{\text{B}}(k) = \min_{r \in [a_{\text{min},B},\, a_{\text{max},B}]} r \qquad s.t.\,(2) \tag{4}$$

to re-enter the right lane behind Vehicle A as fast as possible. If the manoeuvre for re-entering the right lane behind Vehicle A is not completed within the following 4 seconds, the overcoming manoeuvre will be resumed, so that the acceleration switches back to (3). This alternating mechanism proceeds iteratively until Vehicle B re-enter the right lane.

The interaction attack we wish to simulate aims at controlling Vehicle A so that Vehicle B cannot re-enter the right lane, thus exposing it to the risk of crashing against incoming vehicles. To do so, we design a controller for the acceleration of Vehicle A as

$$a_{\text{A}}(k) = \widehat{a_{\text{B}}}(k) + K_{\text{p}}\big(x_{\text{B}}(k) - x_{\text{A}}(k)\big) + K_{\text{v}}\big(v_{\text{B}}(k) - v_{\text{A}}(k)\big) \tag{5}$$

where $\widehat{a_\mathrm{B}}$ is an estimate of the acceleration of Vehicle B, better detailed below, and K_p and K_v are the parameters of the controller.

To show how forecasting or knowing the interaction behaviour of Vehicle B can change the outcome of the attack, we present two cases:

1) case 1 – interaction-unaware attacks: we suppose that the mechanism that governs the acceleration a_B is unknown. We also assume that no smart predictor is implemented to generate a likely estimate of a_B. Instead, we consider only a "naïve" estimate based on the last observed acceleration of Vehicle B, that is

$$\widehat{a_\mathrm{B}}(k) = a_\mathrm{B}(k - 1) \tag{6}$$

2) case 2 – interaction-aware attacks: we suppose that the alternating mechanism that governs the acceleration of Vehicle B is known or forecasted through a data-driven predictor. This means that, ideally, the switching time of a_B (i.e., every 4 seconds) is known. Therefore, we set

$$\widehat{a_\mathrm{B}}(k) = 0 \tag{7}$$

in every 1 second timespan preceding each switch of Vehicle B, to anticipate it. In any other time span, instead, we keep on adopting (6).

The implementation details presented so far, and the relevance of the case study, deserve now a more-in-depth explanation. For many reasons, the proposed case study is nontrivial. Indeed, designing a control logic for our interaction attack to succeed is not straightforward.

First, the behaviour of Vehicle B is dynamic, meaning that it changes in time. To use technical jargon from classical control theory, a_B acts as a disturbance in the control problem of Vehicle A – i.e., a_B is a quantity that cannot be directly controlled. Also, at each timestep, we do not know the value of acceleration that Vehicle B is going to execute. This explains why the "naïve" estimator (6) is based on the observed value at time $(k - 1)$, and not at time k. In (7), instead, we suppose that the acceleration of Vehicle B is known or predicted, which renders it a known external input and no more a disturbance.

Second, Vehicle B is designed by choice as more powerful, that is, to allow higher minimum and maximum jerk than Vehicle A. This means that, without knowing or forecasting the behaviour of Vehicle B beforehand (case 1), it would not be possible to prevent it from re-entering on the right lane. Instead, as in case 2, we aim to show that the interaction attack can be completed by anticipating this information.

The results of the case study are now presented. The characteristics of the two CAVs are reported in Table 1. The parameters of the control problem are instead $\Delta t = 0.01$ s, $K_\mathrm{p} = 0.25$, and $K_\mathrm{v} = 0.5$.

Table 1. Parameters of the case study

L	$\delta_{\text{max, i}}$	$v_{\text{min, i}}$	$v_{\text{max, i}}$	$a_{\text{min, i}}$	$a_{\text{max, i}}$	$j_{\text{min, A}}$	$j_{\text{max, A}}$	$j_{\text{min, B}}$	$j_{\text{max, B}}$
[m]	[rad]	[m/s]	[m/s]	[m/s^2]	[m/s^2]	[m/s^3]	[m/s^3]	[m/s^3]	[m/s^3]
2.9	π/6	10	30	-7	4	-10	5	-11	6

We consider the interaction attack failed when Vehicle B manages to gain a longitudinal distance from Vehicle A of more than 6 meters, either in front of or behind it, within 20 seconds from the beginning of the simulation. The 20-seconds time span represents the time before a third vehicle is expected to arrive from the opposite direction (left lane).

The throttle of the CAVs in the two cases is reported in Figure 2. It is possible to observe that in case 1 (interaction-unaware attack) the throttle of Vehicle A follows the throttle of Vehicle B with a delay, as a consequence of (6). In case 2 (interaction-aware attack), instead, using (7) it is possible to anticipate the next throttle commands of Vehicle B – i.e., the switching behaviour – and so to deceitfully manipulate the manoeuvres of Vehicle B. As a consequence, Figure 3 shows that Vehicle B manages to escape in case 1, but not in case 2. Through interaction-aware attacks of this sort, then, deceitful CAVs would be capable of exerting indirect control through interactions even over more powerful CAVs, and put them in safety critical traffic situations.

Fig. 2. Throttle commands in the two cases

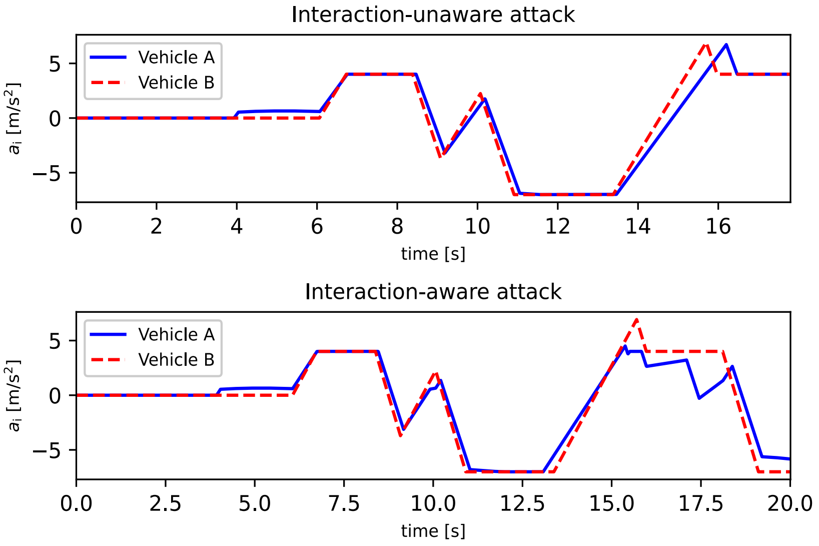
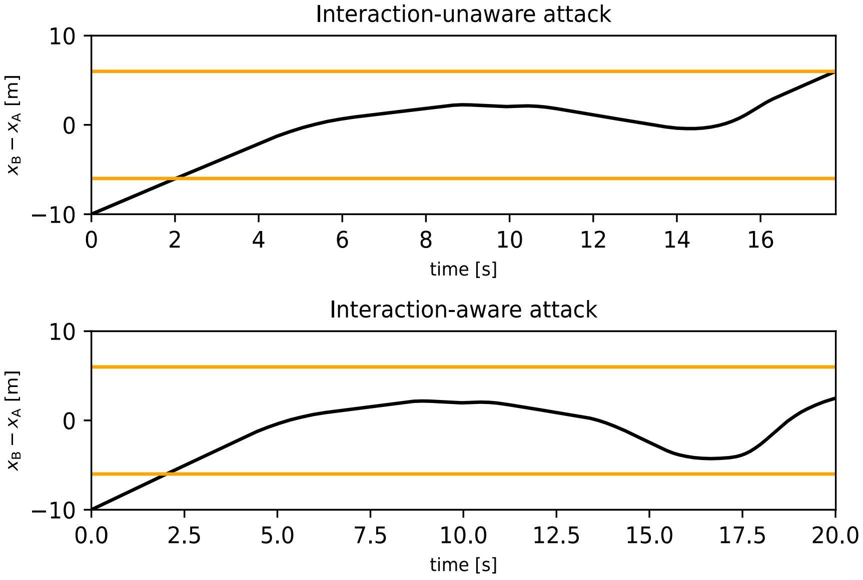


Fig. 3. Relative longitudinal position between the two CAVs in the two cases. The objective of the interaction attack, which starts approximately after 5 seconds, is to keep the relative longitudinal position between the CAVs (black lines) within the 6 meters threshold (orange lines).



6 Interaction Attacks as Deceitful CAV Behaviour

As the previous section shows, the possibility of interaction attacks performed by deceitful CAVs at the expenses of other automated vehicles cannot be ruled out. On this account, we claim that their study should be included into the inquiry on driving automation and deception to which this book is dedicated. In this last section, we draw on the framework advanced in [3] to provide some preliminary observations on interaction attacks as a form of deceitful CAV behaviour.

Nikitas and colleagues [3] have proposed "the term *deceitful* CAV to encompass a vehicle that is deliberately trying to deceive the smart traffic network based on an ulterior motive". Furthermore, they clarify that CAVs could be categorized as deceitful when they "hide", "manipulate", and "falsify deliberately information about their travel intentions, path choices and real-time driving decisions to get an unfair advantage over other vehicles". Since interaction attacks imply a deliberate disguise of real-time driving decisions aimed at putting the attacked vehicle in danger, CAVs executing them would fit well into the category. More precisely, interaction attacks might belong to the class of deceitful behaviour defined as "Target Conditioning" – i.e., "conditioning the data recipient through repetitive behaviour to ensure they believe a normal course of action is being prepared, when in fact a different course of action is being prepared" [3: 4].

As the last quotation suggests, a peculiar dimension of machine deception emerges here – one that has to do with disguising ulterior motives and hidden agendas. Our case displays systems competing with each other while occupying the same space. In a sense, they are playing by the same rules. However, one system deceives the other by pretending to be following an ordinary goal by means of ordinary traffic manoeuvres, while actually pursuing an unordinary, malicious goal through fakes. Arguably, CAVs are built under the presupposition that other CAVs will pursue the goal of getting to their destinations safely, i.e., while avoiding collisions. The attacker exploits this presupposition, thanks to which its moves are not recognised as part of an attacking strategy, but as part of an ordinary driving strategy. When interacting with the attacker, the attacked vehicle does not compute that the attacker's moves might be actually fakes aimed at putting it in a position of danger. Malicious intent is disguised as ordinary driving interaction until the deceived is trapped. Therefore, as Wiener prefigured, machine deception can also pertain to the execution of fakes and feints through which malicious intents are dissimulated, control is indirectly exerted, and attackers lead targets into unpredicted positions of danger.

In their categorisation of deceitful CAV behaviour, Nikitas and colleagues make large use of terms that belong to the semantic field of human deception. "Deliberately", "ulterior motive", "hide", "manipulate", "falsify", "intentions", "believe" are all concepts that primarily belong to the life of the mind, and only analogously apply to the functioning of machines. This linguistic process, that has been defined as the *game of semantic extension* [27] and characterises how artificial agents are made sense of, is not void of risks. As already noted, the risk of projecting mental characteristic onto machines is worrisome and must be attentively curtailed. The duplicity exhibited by the deceitful CAV disguising its actual intention through ordinary interactions should not be conflate beyond its rhetorical usefulness. The attacking CAV just function as it is supposed to, as also the attacked CAV does. The duplicity and the deception *as we commonly denote it* lie in the intent embodied in the deceitful CAVs – the intent of the human attackers deploying it. And those deceived – in the moral sense of the expression – are the humans who built and use CAVs expecting them to be safe and resilient. The attacked CAV can only be said to be deceived *to the extent that* an external agent has intentionally and successfully caused a divergence between the data on which the system bases its operations and the actual state of affairs. The moral, social, and legal ramification of such technical divergence belong to the sphere of human deception.

This clarification is useful to start sketching strategies for counteracting interaction attacks. The moral and the technical components are separated and must be addressed accordingly. On the technical side, measures to minimise divergences between system data and actual states of affairs must be developed. On the moral side, design and development of deceitful CAVs must be dealt with by focusing not on the involved driving systems, but on the humans who, through CAV technologies, deceive and are deceived in the ordinary sense of the world. Mixing the two meanings that the words belonging to the semantics of deception assume in this instance would risk confusing the technical and moral significance of the phenomenon and, ultimately, would lead to ineffective responses.

In light of the above, there are sound reasons to claim that interaction attacks should be added to the list of possible deceitful CAV behaviours and included in the set of risks that cybersecurity should minimize. The prospect of malicious agents manipulating system behaviour by exploiting the features of driving automation might have significant effects both on physical integrity and acceptance. Therefore, future research should be devoted to develop a more fine-grained understanding of this form of attacks and possible countermeasures to minimise the related risks, both of technical and social nature.

References

1. Fossa, F.: Ethics of Driving Automation. Artificial Agency and Human Values. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-22982-4
2. Paparusso, L., Fossa, F., Braghin, F.: Gaming the Driving System. On Interaction Attacks against Connected and Automated Vehicles. In: Fossa, F., Cheli F. (eds.) Connected and Automated Vehicles. Integrating Engineering and Ethics, forthcoming. Springer, Cham (2023).
3. Nikitas, A., Parkinson, S., Vallati, M.: The Deceitful Connected and Automated Vehicle: Defining the concept, contextualizing its dimensions and proposing mitigation policies. Transport Policy 122, 1-10 (2022). https://doi.org/10.1016/j.tranpol.2022.04.011
4. Eugensson, A., Brännström, M., Frasher, D., Rothoff, M., Solyom, S., Robertsson, A.: Environmental, Safety, Legal and Societal Implications of Autonomous Driving Systems. In: 23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV): Research Collaboration to Benefit Safety of All Road Users, 13-0467, pp. 1-15 (2014). http://www-esv.nhtsa.dot.gov/Proceedings/23/isv7/main.htm
5. Collingwood, L.: Privacy implications and liability issues of autonomous vehicles. Information & Communications Technology Law 26(1), 32-45 (2017). http://dx.doi.org/10.1080/13600834.2017.1269871
6. Karnouskos, S., Kerschbaum, F.: Privacy and Integrity Considerations in Hyperconnected Autonomous Vehicles. Proceedings of the IEEE 106(1), 160-170 (2018). https://doi.org/10.1109/JPROC.2017.2725339
7. Taeihagh, A., Lim, H. S. M.: Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. Transport Reviews 39(1), 103-128 (2019). https://doi.org/10.1080/01441647.2018.1494640
8. Banerjee, S.: Autonomous vehicles: a review of the ethical, social and economic implications of the AI revolution. International Journal of Intelligent Unmanned Systems 9(4), 302-312 (2021). https://doi.org/10.1108/IJIUS-07-2020-0027
9. Hansson, S.O., Belin, MÅ., Lundgren, B.: Self-Driving Vehicles—an Ethical Overview. Philosophy of Technology 34, 1383-1408 (2021). https://doi.org/10.1007/s13347-021-00464-5
10. Nikitas, A., Njoya, E. T., Dani, S.: Examining the myths of connected and autonomous vehicles: analysing the pathway to a driverless mobility paradigm. International Journal of Automotive Technology and Management 19(1-2), 10-30 (2019). https://dx.doi.org/10.1504/IJATM.2019.098513
11. Liu, N., Nikitas, A., Parkinson, S.: Exploring expert perceptions about the cyber security and privacy of Connected and Autonomous Vehicles: A thematic analysis approach. Transportation Research Part F: Traffic Psychology and Behaviour 75, 66-86 (2020). https://doi.org/10.1016/j.trf.2020.09.019

12. Le, V.H., den Hartog, J., Zannone, N.: Security and privacy for innovative automotive applications: A survey. Computer Communications 132, 17-41 (2018). https://doi.org/10.1016/j.comcom.2018.09.010
13. Lim, H. S. M., Taeihagh, A.: Autonomous Vehicles for Smart and Sustainable Cities: An In-Depth Exploration of Privacy and Cybersecurity Implications. Energies 11(5), 1062 (2018). https://doi.org/10.3390/en11051062
14. Katrakazas, C., Theofilatos, A., Papastefanatos, G., Härri, J., Antoniou, C.: Cyber security and its impact on CAV safety: Overview, policy needs and challenges. In: Milakis, D., Thomopoulos, N., van Wee, B. (eds.) Advances in Transport Policy and Planning, pp. 73-94. Academic Press (2020). https://doi.org/10.1016/bs.atpp.2020.05.001
15. Aliwa, E., Rana, O., Perera, C., Burnap, P.: Cyberattacks and Countermeasures for In-Vehicle Networks. ACM Computing Surveys 54(1), Article 21 (2022). https://doi.org/10.1145/3431233
16. Rizvi, S., Willet, J., Perino, D., Marasco, S., Condo, C.: A Threat to Vehicular Cyber Security and the Urgency for Correction. Procedia Computer Science 114, 100-105 (2017). https://doi.org/10.1016/j.procs.2017.09.021
17. Petit, J., Shladover, S. E.: Potential Cyberattacks on Automated Vehicles. IEEE Transactions on Intelligent Transportation Systems 16(2), 546-556, (2015). https://doi.org/10.1109/TITS.2014.2342271
18. Lim, B.S., Keoh, S.L., Thing, V.L.: Autonomous vehicle ultrasonic sensor vulnerability and impact assessment. In: 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), 5-8 February 2018, Singapore, pp. 231-236. IEEE Press, Piscataway (2018). https://doi.org/10.1109/WF-IoT.2018.8355132
19. Eykholt, K., Evtimov, I., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., Song, D.X.: Robust Physical-World Attacks on Deep Learning Models. arXiv: Cryptography and Security (2017). https://arxiv.org/pdf/1707.08945.pdf
20. Wiener, N.: Cybernetics or, Communication and Control in the Animal and in the Machine. 2nd Edition. Martino Publishing, Mansfield Center (2013)
21. Kipling, R.: Rikki Tikki Tavi. Wilder Publications, Radford (2021).
22. Rosenblueth, A., Wiener, N., Bigelow, J.: Behavior, Purpose, and Teleology. Philosophy of Science 10(1), 18-24 (1943)
23. Färber, B.: Communication and Communication Problems Between Autonomous Vehicles and Human Drivers. In: Maurer, M., Gerdes, J., Lenz, B., Winner, H. (eds.) Autonomous Driving, pp. 125-144. Springer, Berlin-Heidelberg (2016). https://doi.org/10.1007/978-3-662-48847-8_7
24. Sparrow, R., Howard, M.: When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. Transportation Research Part C: Emerging Technologies 80, 206-215 (2017). https://doi.org/10.1016/j.trc.2017.04.014
25. Loh, W., Misselhorn, C.: Autonomous Driving and Perverse Incentives. Philosophy of Technology 32, 575-590 (2019). https://doi.org/10.1007/s13347-018-0322-6
26. Millard-Ball, A.: Pedestrians, Autonomous Vehicles, and Cities. Journal of Planning Education and Research 38(1), 6-12 (2018). https://doi.org/10.1177/0739456X16675674
27. Fossa, F.: Artificial Agency and the Game of Semantic Extension. Interdisciplinary Science Reviews 46, 440-457 (2021). https://doi.org/10.1080/03080188.2020.1868684