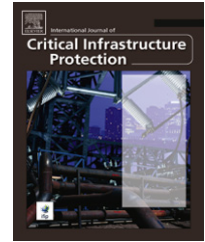


Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SciVerse ScienceDirect

journal homepage: [www.elsevier.com/locate/ijcip](http://www.elsevier.com/locate/ijcip)

# An exploration of defensive deception in industrial communication networks

Julian L. Rrushi

Faculty of Computer Science, University of New Brunswick, 550 Windsor St., Fredericton, New Brunswick E3B 5A3, Canada

## ARTICLE INFO

### Article history:

Received 19 May 2010

Received in revised form

23 March 2011

Accepted 18 June 2011

Published online 30 June 2011

### Keywords:

Industrial informatics

Military deception

Intrusion detection

Signal detection theory

## ABSTRACT

Process control networks constitute a vantage point for computer network attacks on electrical power infrastructures such as power plants and electrical substations. Consequently those networks represent a critical point of network defense in power grid computer networks. In this paper we discuss research that draws on military deception to conduct a cognitive hacking into the attacker's mind at the process control network level. This research enables the defender to influence the attacker's target selection process, and thus pilot it towards simulated physical processes and equipment. A hijacked target selection process causes the attacker to generate specific network traffic that makes a significant contribution to the detection of the ongoing network intrusion. Our cognitive hacking approach is based on displays created via simulation of the appearance of physical processes and equipment. The main counter attack vectors employed consist of emission of deceptive network traffic and exploitation of information conversion as means of concealing deceptive simulation. We have implemented this research as a small proof of concept prototype, and thus in the paper we also discuss an analysis of its deception effects via application of signal detection theory.

© 2011 Elsevier B.V. All rights reserved.

## 1. Introduction

Malicious computer network intrusions in power grid computer networks are a concrete threat for physical destruction of electrical power infrastructures such as power plants and electrical substations [1]. That theory was validated through ethical research conducted at the US DoE Idaho National Laboratory (INL) within a project named Aurora [2]. INL researchers conducted an experimental computer network attack on the replica of a process control system that is typically used to monitor and control an electrical power generator, which is a common equipment component to power plants. The final outcome of the experiment was a violent physical destruction of the electrical power generator in question. The network reconnaissance that precedes such computer network attacks includes identifying the type and technical

characteristics of physical equipment monitored and controlled by process control systems, and discovery of some of the technical details behind the interaction between the two.

That specific network reconnaissance is conducted at the process control network level. Thus, process control networks constitute a vantage point for computer network attacks on electrical power infrastructures. In this paper we discuss a defensive deception approach, namely a novel application of military deception (MILDEC) [3], which aims at influencing the attacker's target selection process at the process control network level. We coded this approach with the name of mirage theory. The approach exploits the network reconnaissance process as a practical means of penetrating the attacker's mind with technical information and indicators such as to hijack the attacker's target selection process in a way that facilitates detection of the ongoing

E-mail address: [jrrushi@unb.ca](mailto:jrrushi@unb.ca).

intrusion. Our research was inspired by a lesson that we drew from history, namely Operation Fortitude South conducted during the second world war [4].

Operation Fortitude South was a strategic plan that preceded the allied invasion of German occupied territory of France. The ultimate objective of Operation Fortitude South was to deceive the German military command into believing that the allies would attack from Pas de Calais rather than from Normandy. Operation Fortitude South comprised creation and deployment of a special electronic unit known as the 5th wireless group along with large intelligence operations such as espionage and controlled leaks of information through diplomatic channels. The 5th wireless group used some newly developed radio transmitters that ran pre-programmed and especially written scripts to generate radio communications. Those radio communications consisted of conversations that are typical to military assault operations.

The German military in occupied France had few aerial reconnaissance capabilities left by the time Operation Fortitude South was conducted. Eavesdropping on radio communications was the principal mechanism that they could use to determine movements of allied troops. Operation Fortitude South was highly successful to a degree that, upon recommendation of the German military command, Adolf Hitler concentrated a large amount of military capability, including Panzer tank units, in Pas de Calais. In mirage theory we exploit similar concepts as Operation Fortitude South, namely the attacker's reliance on analysis of intercepted network traffic to derive the presence and characteristics of physical targets in electrical power infrastructures, and the attacker's lack of means to verify that intercepted network traffic is indeed due to occurrence of existing physical processes and operation of existing physical equipment in electrical power infrastructures.

## 2. Related research

In [5,6], Rowe and Rothstein explore deception techniques drawn from conventional warfare for improving the security of computer systems and networks. The authors analyze Operation Mincemeat [7] in order to illustrate a set of principles and mechanisms that are used for an effective tactical deception in conventional warfare. Operation Mincemeat is a historical military operation that took place during the second world war. The authors then evaluate the applicability of those principles and mechanisms to the invention of defensive deceptive capabilities for computer systems and networks. Mirage theory moves along the line of Rowe and Rothstein's research as it is a direct application of MILDEC to the defense of process control systems and networks. Similar to the work of Rowe and Rothstein, mirage theory was devised upon analysis of a historical conventional warfare operation.

Mirage theory has a few features in common, to some extent, with honeypots, i.e. closely monitored computer system resources that serve as network decoys [8,9]. Those features comprise distraction of attackers from valuable attack targets and leverage of deception for intrusion detection [10,11]. Nevertheless, mirage theory is fundamentally different. Honeypots are totally passive and just stand by to receive

network connections from attackers or autonomous attack agents such as worms or bots. Thus, honeypots do not have the host and network activity that is commonly found in production computer systems. Low interaction honeypots like Honeyd [12] respond to network probes, but they do not allow system access as their vulnerable services are emulated via specific scripts.

Because of that reason, a low interaction honeypot is easily detected after an initial exploitation attempt. High interaction honeypots respond to network probes and also allow the attacker or autonomous attack agent to obtain system access. Nevertheless, monitoring system events in the compromised honeypot along with network traffic that is seen at its network interface card reveals the lack of realism in its host and network activity. Mirage theory is exactly the opposite of honeypots in that regard, in the sense that the main strength of mirage theory lies in host and network activity that is commonly found in process control systems and networks in production. Honeypots employ real or emulated services that are more vulnerable than their production counterpart in order to lure attackers. Mirage theory does the opposite, namely it makes deceptive process control systems and networks and simulated physical processes and equipment as much undistinguishable from their production counterparts as possible.

The deception capabilities of honeypots are placed within the boundaries of a computer system, and hence fall within network access visibility. Unlike honeypots, mirage theory develops deceptive capabilities at a layer that is not reachable through network access to a compromised process control network. Rowe and Rothstein in [5,6] indicate that honeypots are not in line with an important principle of conventional warfare, namely that deception should be integrated with genuine operations. The authors argue that deceptive tactics are more effective on real systems. In fact Holz and Raynal in [13] provide several techniques that an attacker could use to detect honeypots. Those techniques detect technical details that are characteristic of virtual execution environments, which in turn are commonly used for implementing high interaction honeypots.

Mirage theory employs genuine process control systems and networks, which are deployed and configured such as to smoothly monitor and control an existing electrical power infrastructure. The defender could utilize cheap hardware for those process control systems and networks as they are not subject to the same strict reliability and physical robustness requirements as their production counterpart. In [14], Yuill et al. discuss a deception-based intrusion detection approach that leverages the concept of honeyfiles, which are similar to some degree to deceptive program variables in cyber-physical mappings employed by mirage theory. A cyber-physical mapping is a one-to-one correspondence between control application variables and physical process parameters or parameters that characterize the operation of physical equipment. Those variables hold I/O values in the random access memory (RAM) of process control systems.

Honeyfiles are bait files that are intended for an attacker to access. A honeyfile is constructed such as the computer system in which that honeyfile resides will generate intrusion alerts if the honeyfile is accessed. Honeyfiles are intended

to be no different than other ordinary files in the file system of a computer system. For an attacker to detect a honeypot, that attacker has to open the honeypot. Clearly that action will result in detection of the ongoing intrusion. While a file is mapped to regions of secondary storage by the operating system, a deceptive variable in mirage theory is mapped to a parameter related to a physical process or equipment, which in fact are all simulated. Mirage theory employs deceptive network packets to make deceptive variables appear no different than their genuine counterpart. For an attacker to detect a deceptive variable, that attacker has to access the deceptive variable either locally or over a process control network. Accessing the deceptive variable will cause detection of the ongoing intrusion.

In [15], Rowe addresses the problem of logical consistency in deception. The author explores automated methods which track assertions that have been made up to a certain point in time along with their effects. Those automated methods thereafter identify possible consistent deceptive actions that may be conducted next in order. In mirage theory we address the same problem. We do so in a way that differs from the approach followed by Rowe in [15] due to our different levels of intervention. Rowe works mainly at the operating system level, while in mirage theory we focus mainly on simulation of physical processes and equipment to ensure logical consistency in deception. That deceptive simulation aims at feeding an attacker with a consistent view of the internal dynamics of physical processes and equipment at any point in time.

### 3. The mirage approach

#### 3.1. Relation to conventional warfare theories

MILDEC forms the basis of mirage theory. MILDEC is defined as those actions executed to deliberately mislead adversary decision makers as to friendly military capabilities, intentions, and operations, thereby causing the adversary to take specific actions or inactions that will contribute to the accomplishment of the friendly mission [3]. We can define mirage theory as actions that are devised to deliberately mislead an attacker as to electrical power infrastructures, thereby causing the attacker to take specific actions that will contribute to detection of the ongoing intrusion. Deception means in MILDEC are grouped into three categories, namely physical means, technical means, and administrative means [16]. Examples of physical means include dummy and decoy equipment and devices, tactical actions, movement of military forces, etc.

Examples of technical means include emission of chemical or biological odors, emission of radiation, reflection of energy, computers, etc. Examples of administrative means include techniques to convey or deny physical evidence. Mirage theory employs mainly technical deception means, namely emission of deceptive network traffic. Mirage theory relies to a large degree on a MILDEC concept that is referred to as a display. Displays are simulation, disguising, and/or portrayal of friendly objects, units, or capabilities that may not exist,

but are made to appear so. In that regard, mirage theory employs computer clusters to simulate the presence of physical processes and equipment. The ultimate objective is what mirage theory has in common with cognitive hacking, perception management, and reflexive control theory. Cognitive hacking is basically manipulation of the perception of technology users [17].

Perception management is comprised of actions that convey and/or deny selected information and indicators to foreign audiences in order to influence them such as to result in behaviors favorable to the originator's objectives [18]. Reflexive control is a warfare theory that has been studied in the former Soviet Union and later on in Russia for a very long time. Reflexive control theory is comprised of methods for conveying especially prepared information to a subject in order to incline that subject to voluntarily make a predetermined decision [19]. Mirage theory seeks to exploit the attacker's mind, namely the attacker's perception of an electrical power infrastructure. Mirage theory actively conveys information and indicators to an attacker for the purpose of influencing that attacker's target selection process such as to hijack it towards displays of physical processes and equipment.

#### 3.2. Display creation

The cognitive hacking approach that we take in mirage theory is based on displays that faithfully mimic the appearance of an electrical power infrastructure. In this research we create those displays via real-time computer simulation of physical processes and equipment. A viable simulation technique for creating those displays is continuous simulation [20], as physical processes and equipment in electrical power infrastructure are mostly continuous in nature. We develop models of physical processes and equipment in Matlab Simulink [21]. Those models comprise ordinary or partial differential equations that represent the internal dynamics of our physical processes and equipment of reference at any point in time. We then convert those models into C language code via the Real-Time Workshop tool [22]. The deceptive real-time simulation of our physical processes and equipment of reference is conducted by running that C code on a cluster of personal computers.

We conducted this work with specific reference to a nuclear power plant that follows a boiling water reactor design. We practically conducted deceptive simulation of an alternating current (AC) induction motor that drives a water pump, which both are common equipment components to such power plants. The C code that corresponds to the Simulink models interfaces with the execution environment via environment variables such as to allow a computer program to read status parameters that characterize the state of the simulated process or equipment, and also write parameters that change the state of the simulated process or equipment. Thus, that specific I/O interaction via reading and writing of environment variables allows for computer program emulation of sensors and actuators commonly found in real electrical power infrastructures. For example, in our prototype a computer program could read environment variables that represent the rotational speed of the simulated

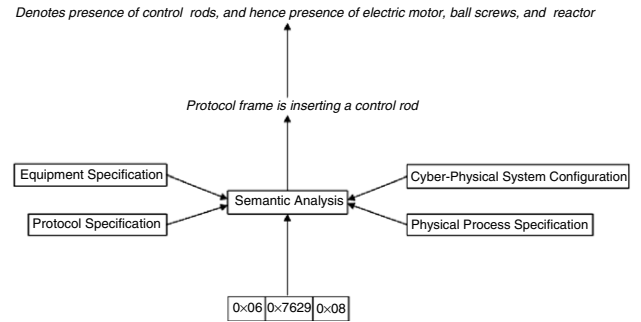
AC induction motor and the injection rate of the associated simulated water pump, respectively.

Similarly, a computer program could set the value of an environment variable that represents the applied voltage frequency of the simulated AC induction motor. Applied voltage frequency is a physical parameter that controls the actual rotational speed of an AC induction motor, and hence the injection rate of the associated water pump. Computer program emulation of sensors and actuators in our deceptive simulation allows for interactivity with the attacker. The defender could have the attacker generate malicious network packets that aim at disrupting physical processes and equipment, and thereafter verify the possible negative impact of those packets. That is useful in the case the defender is interested in allowing the attacker to make some progress in order to extract a better characterization of the ongoing computer network attack. It may also become necessary for the defender to allow transmission of more than a single malicious network packet so that to gain sufficient confidence that a computer network attack is actually taking place.

In mirage theory we connect the cluster of personal computers that conducts deceptive simulation to genuine process control systems over a genuine process control network. Control applications running on those process control systems estimate the state of the simulated electrical power infrastructure by reading values of specific parameters from computer program emulated sensors. Those control applications change the state of the simulated electrical power infrastructure by setting values of specific parameters via computer program emulated actuators. Monitoring and control of the simulated electrical power infrastructure results in generation of network traffic in the process control network as if the simulated electrical power infrastructure were real. As we were conducting this research, we were expecting imperfections in that network traffic. The reason is that we solve the differential equations of our Simulink models via numerical analysis.

Numerical analysis generally produces numerically approximated solutions of differential equations. In practice, though, we did not see approximate solutions of the differential equations of our Simulink models as an anomaly as those solutions lie within an acceptable degree of accuracy range. The network traffic that is generated by monitoring and control of an existing electrical power infrastructure does not reflect any absolute perfection. A possible instance of a source of imperfection in that network traffic is the data conversion process in sensors and actuators. There are no ideal analog-to-digital converters. Analog to digital conversion of data is characterized by unavoidable errors such as quantization errors, aperture errors, nonlinearity errors when applicable, etc. Similarly, digital-to-analog conversion of data is not ideal either. The errors that apply to the data conversion process are mostly random.

If the attacker were to analyze network traffic in a process control network to determine whether the target electrical power infrastructure is existing or deceptively simulated, the challenge that the attacker would face consists in how to differentiate between randomly imperfect views of the target electrical power infrastructure. From a network access perspective, the existence of physical processes and physical



**Fig. 1 – A network packet payload that is indicative of the existence of physical equipment and a physical process.**

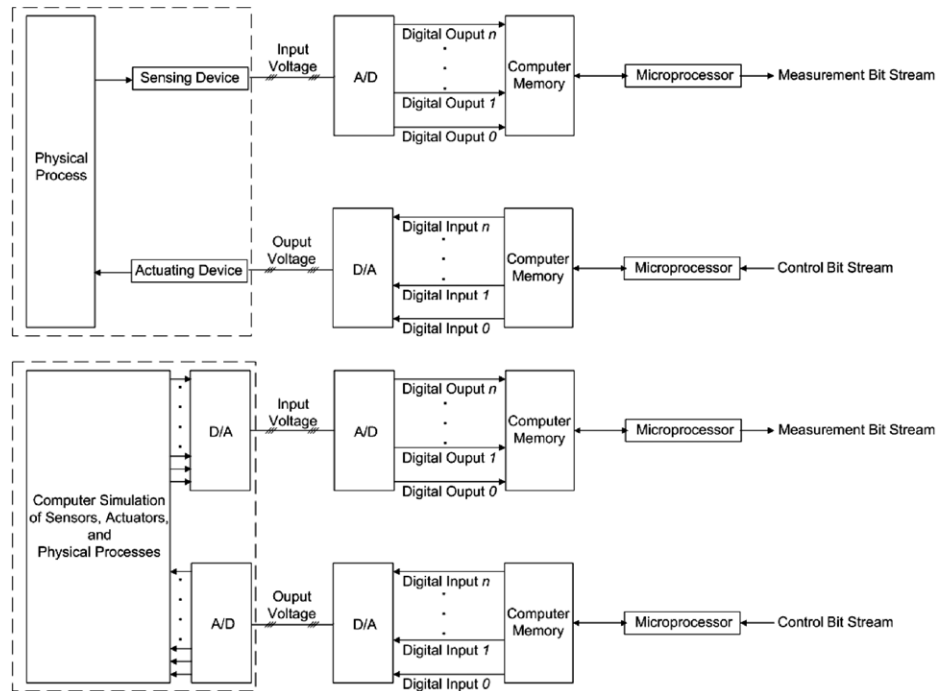
equipment is derived from network packets that flow over a process control network. For example, a network packet such as the one depicted in Fig. 1 sniffed from the process control network of a nuclear power plant denotes the presence of an electric motor that generates rotational motion, a ball screw that translates this rotational motion into a linear motion, and a control rod that is inserted or withdrawn via the linear motion in question.

A computer network attack that aims at physical destruction of such physical equipment requires a complete technical profile of that equipment. The attacker constructs that technical profile via analysis of network traffic sniffed from the process control network. In this research we consider the general case in which the attacker attempts physical destruction of an electrical power infrastructure exclusively through computer network attacks. Clearly the attacker can combine computer network attacks with physical world acts. For example, the attacker could obtain the technical profile in question directly from insider threats. The defense coverage of those scenarios lies outside the scope of this research. The display in mirage theory results in generation of network traffic that guides the attacker's analysis of that traffic into construction of technical profiles of physical processes and equipment, which in fact are all simulated. Mirage theory leverages those technical profiles to exploit the attacker's target selection process, and thus deceive the attacker into targeting the display.

### 3.3. Display concealment

In mirage theory we leverage conversion of data from analog-to-digital and vice versa as a physical means of preventing the attacker from gaining network access to the cluster of personal computers that creates displays. In this subsection we treat display concealment in relation to electrical power infrastructures in which data conversion is conducted at edge control systems. An edge control system is a process control system that is directly wired to sensors and/or actuators. Edge control systems are located at the edges of the process control network in very close vicinity to physical equipment. Several electrical power infrastructures employ digital sensors and actuators, which are otherwise known as smart instruments. In those electrical power infrastructures data conversion is conducted at the smart instruments, which are equipped





**Fig. 2 – Boundary between cyber components and physical components in an electrical power infrastructure exploited in mirage theory to camouflage displays.**

with computing resources of their own, i.e. microprocessor, RAM, flash memory, etc.

Our treatment of display concealment applies to those electrical power infrastructures similarly, with the only difference being that the physical means of concealing the displays lie in smart instruments rather than in edge control systems. The interactions between sensors or actuators and edge control systems take place via application of electrical signals with certain characteristics. That process is illustrated in the top part of Fig. 2. In a typical sensing activity, sensors, i.e. transducers, measure physical phenomena and hence generate analog data, i.e. voltages or currents, proportional to the measured value. For example, incore detectors in a nuclear reactor measure neutron flux. Those incore detectors apply electrical signals that are proportional to neutron population in the reactor core.

The neutron flux measurements conveyed by those electrical signals are processed by edge control systems, which together form a neutron monitoring system. An edge control system is equipped with analog-to-digital conversion integrated circuits [23], which periodically sample and convert those electrical signals into discrete numerical values. Edge control systems actuate physical equipment also by applying electrical signals. For example, an edge control system may set the actual rotational speed of an AC induction motor by controlling the applied voltage frequency. An edge control system is equipped with digital-to-analog conversion integrated circuits, which convert discrete numerical values into electrical signals. In mirage theory we perceive analog-to-digital and digital-to-analog conversion integrated circuits as a boundary between process control systems and networks and physical processes and equipment.

A computer network intrusion enables the attacker to access the process control systems in the compromised process control network. Nevertheless, a computer network intrusion by no means will enable an attacker to virtually move beyond the analog-to-digital and digital-to-analog conversion integrated circuits. Intuitively we position the cluster of personal computers that creates displays behind the boundary, and hence use that boundary to conceal displays. The attacker cannot verify whether input electrical signals are indeed generated by existing sensors, nor can the attacker verify whether output electrical signals indeed reach an existing actuator.

The bottom part of Fig. 2 illustrates how the cluster of personal computers that creates displays correlates and interacts with the process control systems in the process control network. We use a computer program that we wrote in the C language to periodically read status parameters from environment variables which represent the state of the simulated process or equipment. That computer program emulated sensor runs on one of the personal computers in the cluster of computers that conducts the deceptive simulation, which in this research we refer to as the interface computer. The interface computer is wired with analog-to-digital and digital-to-analog conversion integrated circuits. The measurement values read by the computer program emulated sensor are passed to the digital-to-analog conversion integrated circuits, which generate the electrical signals that correspond to those measurement values.

Those electrical signals are received by the analog-to-digital conversion integrated circuits of an edge control system as if the electrical signals in question were generated by an existing sensor. The edge control system then converts those electrical signals into discrete numerical values in

the form of a measurement bit stream, which is processed locally by control applications. The edge control system may propagate the measurement bit stream over the process control network. Possible destinations for the measurement bit stream include one or more upper layer process control systems in the process control network and the human machine interface (HMI) of a system operator. Since it is only after the analog-to-digital conversion that the measurement values in question become accessible to the attacker at the process control network, creation of the display is transparent to the attacker.

A control bit stream is generated by control applications running on the edge control system in order to drive an actuator, which in turn changes the state of a physical process or equipment. The edge control system uses its own digital-to-analog conversion integrated circuits to convert the control bit stream into electrical signals as if it were to control an existing physical process or equipment. In a real electrical power infrastructure those electrical signals are received by an existing actuator, which processes them as they are. In mirage theory we employ the analog-to-digital conversion integrated circuits of the interface computer to receive those electrical signals and convert them into discrete numerical values thereafter. We use a computer program that we wrote in the C language to read and process those discrete numerical values.

That computer program emulates the functioning of a real actuator and runs on the interface computer. Such a computer program emulated actuator changes the state of the simulated processes and equipment by assigning the discrete numerical values in question to specific environment variables, which in turn are read by the ongoing execution of the C code that corresponds to the Simulink models. This way our deceptive simulation recreates the effects that the original electrical signals applied by the edge control system would have had on existing physical processes and equipment in an electrical power infrastructure. After the point in which the control bit stream is converted into electrical signals by the digital-to-analog conversion integrated circuits of the edge control system, that control bit stream is no longer accessible from the process control network.

Thus, the attacker entirely loses visibility on the control bit stream after that point. Consequently, the attacker has no means of telling whether the electrical signals that correspond to the control bit stream are indeed received by an existing actuator. With that result, creation of the display is, again, transparent to the attacker.

### 3.4. Detecting network intrusions

The approach that we take in mirage theory to detect a network intrusion consists in checking the environment variables which represent physical parameters of the simulated electrical power infrastructure for deviations from safe values. We conduct those checks programmatically at the cluster of personal computers that creates displays. That approach is similar to the intrusion detection technique proposed by Cárdenas et al. in [24]. Cardenas et al. model a physical system as a linear dynamical system, and thus

propose to use such model to determine the effects of network packets on physical parameters of that physical system. Both our approach to detecting an ongoing intrusion into the process control network and the proposal of Cárdenas et al. leverage the fact that computer network attacks exhibit an abnormal behavior of the target physical system, i.e. have negative effects on physical parameters of that physical system.

Our work differs from the proposal of Cárdenas et al. in that we allow possibly malicious network packets to hit their target, while Cárdenas et al. propose to use sequential detection theory to determine whether or not network packets will have negative effects on physical parameters before those network packets are processed by a process control system. We are in the conditions of doing so as in our case the physical processes and equipment are all simulated. We sniff and store network packets that flow over the process control network for further forensics processing. We conduct network forensics that exploits complex causality relations that hold in an electrical power infrastructure in order to recognize non-self network traffic, i.e. network traffic that was not generated by the employment of mirage theory. That network forensic approach lies outside the scope of this paper, and consequently we leave it to a separate research paper.

## 4. Analysis of deception effects

### 4.1. Attack–defense model and testbed overview

We devised and experimentally emulated an attack–defense model in a testbed in order to support our analysis of the deception effects of mirage theory. In that model an attacker attempts an intentional loss of cooling accident on a power plant. In the power plant of reference, a number of water pumps feed water into the reactor core. The injected water picks up the heat produced by nuclear fission, and thereafter is transformed into steam. The produced steam is then directed through pipes to spin the shaft of a turbine that is connected to an electrical power generator. In addition to being transformed into steam, the injected water also serves to cool nuclear fuel in the reactor. The water pumps that feed water into the reactor core are driven by electric motors of type AC induction motors. The AC induction motors in turn are controlled and monitored by programmable logic controllers (PLCs), which as other types of process control systems in a power plant are known with the common term of instrumentation and control (I&C) systems.

Each one of those PLCs uses a continuous sensor, namely a battery powered stroboscopic tachometer, to measure the rotational speed of the AC induction motor under its control. The PLCs are part of a distributed control system (DCS) whose process data communications are conducted over the ModBus TCP protocol. A larger view of the testbed is depicted in Fig. 3. In the attack–defense model the attacker exploits coding vulnerabilities in control applications to penetrate the supervisory level of the DCS from the enterprise network. The attacker then conducts a computer network attack on the PLCs. The objective is to cause physical damage to the AC

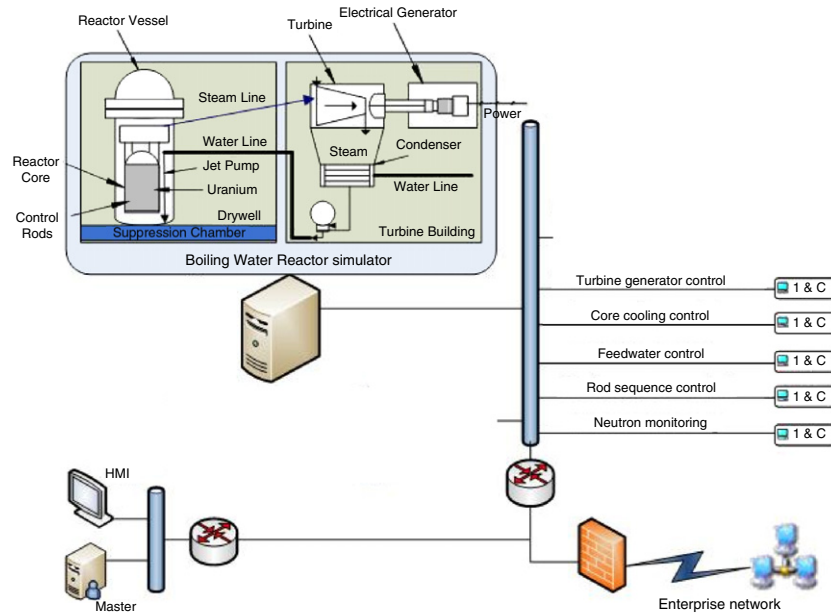


Fig. 3 – A graphical representation of the testbed.

induction motors that the PLCs control, and hence prevent the water pumps from functioning. Dysfunction of water pumps causes a loss of cooling in the reactor.

More specifically, the attacker disrupts a water pump by conducting a network inertial attack on the AC induction motor which drives that water pump. An inertial attack is conducted by speeding up or slowing down heavy equipment at high rates [25]. It is reported to have potential for forcing heavy equipment to fail as heavy equipment is not tolerant to abrupt speed changes [25]. For the attacker to affect the operation of any physical equipment controlled by a PLC, the attacker has to preliminarily identify that part of a cyber-physical mapping which relates control application variables stored in the RAM of the PLC to physical parameters that characterize the operation of the physical equipment in question. This is because the attacker can affect those physical parameters only by modifying the corresponding program variables.

Thus, in the attack-defense model the attacker is required to discover the ModBus address of the program variable that is mapped to applied voltage frequency for being able to conduct any computer network attack that disrupts the target water pump. As we wrote earlier in this paper, applied voltage frequency is a physical parameter that controls the actual rotational speed of an AC induction motor. In the attack-defense model the reconnaissance conducted by the attacker includes application of the approach described in [26], namely a statistical technique that employs the degree of linear association among program variables as measured by a linear correlation coefficient to identify the program variable of interest, and hence discover the corresponding ModBus address. The defense part of the attack-defense model consists in a deployment and activation of mirage theory to detect and deter a loss of cooling attack.

#### 4.2. Analysis

We now discuss human subject testing that we conducted within this research to analyze the effects of mirage theory on the attacker's decision making process in relation to the attack-defense model described in the previous subsection. We conducted those tests through the lenses of signal detection theory. Signal detection theory is a method to characterize and quantify the ability of a subject to discern between signal and noise. The reader is referred to [27,28] for background knowledge on signal detection theory. The subjects whose decision making was integrated into our attack-defense model for deception analysis purposes consisted of a group of students. Thus, those students were assigned the role of attackers. We employed two data sets for these experiments, which we used to refer to as the genuine data set and the mirage data set, respectively.

The genuine data set comprised packet capture (pcap) files that were created from sniffing the network packets received from or sent by PLCs that controlled existing AC induction motors. The mirage data set was developed by us in the following way: we employed the ModBus master data scanner tool [29] to periodically scan the discrete input variables, coil variables, input register variables, and holding register variables from the RAM of the PLC that controlled displays of an AC induction motor and an associated water pump in the testbed. We did so over a several hour time period. The reader is referred to [30] for a specification of the ModBus protocol. We sniffed with Wireshark [31] the network packets that were generated by that network scanning, and thus created pcap files that formed the mirage data set.

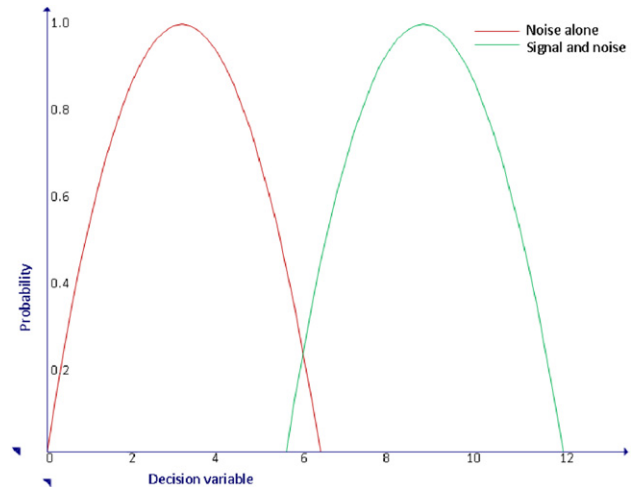
The students were asked to identify the ModBus address of a specific holding register variable that was mapped to applied voltage frequency. That task corresponds to target

selection when considered in relation to a network inertial attack on an AC induction motor which drives a water pump. Thus, we have “signal” when a program variable under consideration is mapped to applied voltage frequency and denotes the presence of an existing AC induction motor, and “noise” otherwise, i.e. the program variable under consideration is not mapped to applied voltage frequency or does not denote presence of an existing AC induction motor. “Internal response” is resembled by a decision variable that quantifies the elements that suggest that signal is present.

The decision variable is related to such measures as the proximity of the linear correlation coefficient estimated for a candidate holding register variable to the linear correlation coefficient that is known to be characteristic for the holding register variable that is mapped to applied voltage frequency, similarity between observed values of a candidate holding register variable and typical values of applied voltage frequency, and the likelihood that the applied voltage frequency applies to an existing AC induction motor. The students extracted the values of ModBus variables from the protocol data units in pcap files. Thus, the data that were analyzed by the students for target selection were in the form shown in Table 1. In Table 1, IR and HR stand for input register variable and holding register variable, respectively. The ModBus address of those scanned variables shown in Table 1 is indicated in square brackets.

Prior to the experiments, we taught the students about the statistical technique described in [26], which in turn required them to study the internals of process control networks, ModBus TCP protocol, PLCs, and AC induction motors. Those students do not equate with expert attackers. Furthermore, the offensive part of our attack-defense model is not a generalization of reconnaissance and attack techniques that might be applied by an expert attacker. Nevertheless, those limitations do not affect the human subject testing part of this research, as our purpose is to analyze the deception effects of mirage theory by observing how it performs against a possible practical and well conducted reconnaissance and attack technique. Thus, although the estimations of deception effects that we achieved from the experiments with our attack-defense model and the group of students do not hold for all reconnaissance and attack techniques and attackers, those estimations shed light on the effectiveness of mirage theory.

We conducted two rounds of trials. In the first round the students were given pcap files from the genuine data set. The students were informed that those pcap files did not contain any mirage data. Each trial consisted in asking the students whether or not a candidate holding register variable was signal, to which they responded yes or no. Firstly, the students were presented noise trials. Thus, none of the candidate holding register variables in those trials were signals. We used the relative-frequency method to estimate the probability distribution of the decision variable for those trials being based on the students' responses. That probability distribution is represented by the left probability of occurrence (POC) curve in Fig. 4. The horizontal axis in Fig. 4 denotes values of the decision variable, while the vertical axis denotes the frequency of the occurrence of each one of those values.



**Fig. 4 – POC curves that represent probability distributions of the decision variable in noise trials (left) and signal trials (right) in the first round of trials.**

The students then were presented signal trials. Thus, all of the candidate holding register variables in those trials were signals. We followed the same steps as with noise trials. The right POC curve in Fig. 4 represents the probability distribution of the decision variable for those signal trials being based on the students' responses. The first round of trials revealed that during the reconnaissance for a loss of cooling attack the students were subject to internal noise, whose most common form observed was that several program variables were found to be linearly associated to the same degree. Internal noise was found also during the application of other optional or complementary reconnaissance techniques. For instance, the students attempted to recognize the holding register variable of interest by comparing the values of reconstructed program variables to typical values of applied voltage frequency.

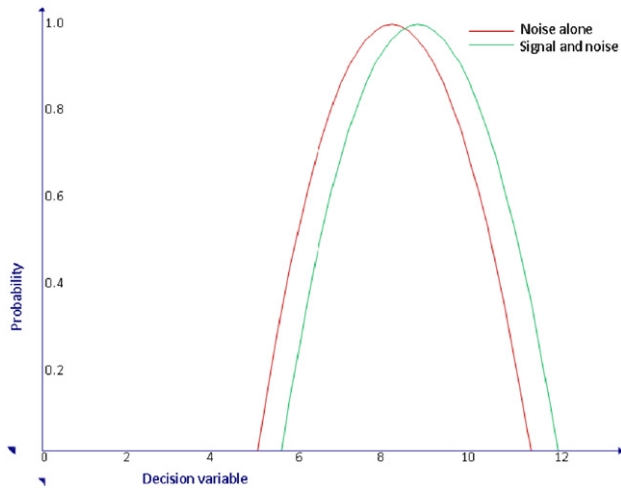
The internal noise in that case was that values of several holding register variables may be typical to applied voltage frequency. Despite the effects of internal noise, this round of trials showed that the students' decision making processes with regard to target selection were subject to a relatively low uncertainty. The POC curves in Fig. 4 show that the signal strength was high and the amount of noise, both external and internal, was low. Consequently the overlap of these two POC curves was small, while their spread was reduced. The discriminability index derived from the separation and spread of the two POC curves in Fig. 4 had a value around  $d' = 5.6$ . Recall from signal detection theory that  $d'$  is an estimate of the signal strength. The formulae that we used to calculate the discriminability index are concisely provided in [32].

With such a high discriminability index an existing AC induction motor was considerably discriminable from its display counterpart. Clearly the students based their responses on the value that the decision variable exhibited at each trial, regardless of whether it was a noise trial or a signal trial. If the decision variable was equal to or greater than a specific criterion, the students responded yes, otherwise their response was no. Each criterion leads to specific hit rates and false alarm rates. We constructed a graph in the form of a receiver operating characteristic (ROC) curve that indicated



**Table 1 – Excerpt from the data analyzed for target selection.**

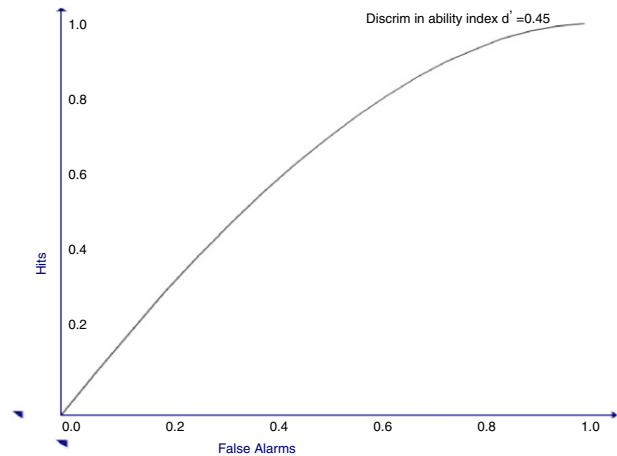
IR[16]	IR[53]	IR[18]	IR[69]	HR[19685]	HR[20008]	HR[18610]
702.5	1884.0	1205.3	685.2	63.9	36.5	42.1
803.8	1977.0	903.9	679.2	55.4	39.2	41.6
901.8	1782.0	1306.9	722.4	55.8	38.3	45.2
904.1	1608.0	1004.8	763.2	67.3	45.8	48.6
1004.7	1884.0	1407.8	735.6	57.8	48.1	46.3
903.1	1977.0	1409.4	796.8	58.1	49.3	51.4
1004.9	1782.0	1408.3	868.8	61.8	51.5	57.4
809.6	1608.0	1598.3	817.2	48.9	58.3	53.1
1208.8	1782.0	1203.9	890.2	38.9	61.8	59.1
803.5	1608.0	957.5	945.6	48.6	47.5	63.8

**Fig. 5 – POC curves that represent probability distributions of the decision variable in noise trials (left) and signal trials (right) in the second round of trials.**

hit rates and false alarm rates for a discriminability index  $d' = 5.6$  as the criterion was shifted from low values towards high values. That ROC curve was plotted with the false alarm rate on the horizontal axis and the hit rate on the vertical axis.

It went up to the upper left corner converging with a straight line that intersected the vertical axis at a value of 100%, and was parallel to the horizontal axis. The ROC curve in question shows that the students' decision making processes with regard to target selection were characterized by a large number of hits and just a few false alarms. In the second round of trials the students were given a mixture of pcap files from the genuine data set and pcap files from the mirage data set. This time we informed the students that the set of pcap files that each one of them was assigned did contain mirage data. We proceeded similarly to the first round of trials. The students were presented noise trials followed by signal trials. The mirage data were part of the noise trials. The POC curves that represent the probability distributions of the decision variable for those noise trials and signal trials being based on the students' responses are depicted in Fig. 5.

Those two POC curves overlap greatly, which indicates that the signal strength decreased as the students fell under the effects of mirage theory. We recalculated the discriminability index and obtained a value of  $d' = 0.45$ , which is considerably lower than the original value of  $d' = 5.6$ . The mirage data

**Fig. 6 – ROC curve developed with results from the second round of trials.**

served as internal noise, which drastically increased the uncertainty to which the students' decision making processes were exposed with regard to target selection. As in signal detection theory, a subject, i.e. an attacker, has little or no control over the internal noise that is emitted during his/her decision making process. The internal noise made an existing AC induction motor hardly discriminable from its display counterpart. The ROC curve that indicates hit rates and false alarm rates for a discriminability index  $d' = 0.45$  as the criterion is shifted from low values towards high values is depicted in Fig. 6. That ROC curve shows that the students' decision making processes with regard to target selection were characterized by a smaller number of hits and a much bigger number of false alarms compared to the first round of trials in which mirage theory was not active.

## 5. Conclusions

In this paper we discussed our research on a defensive deception approach that we coded with the name of mirage theory. Mirage theory aims at a cognitive hacking into the attacker's mind such as to hijack the attacker's target selection process towards displays of physical processes and equipment. A hijacked target selection process is then exploited to unequivocally detect the ongoing intrusion. In the paper we elaborated on how we create displays via

deceptive continuous simulation based on Matlab Simulink models of physical processes and equipment. We also elaborated on how we exploit data conversion as a physical barrier for concealing displays. We implemented this research as a small proof of concept prototype, which enabled us to empirically analyze the deception effects of mirage theory. In the paper we discussed human subject testing, in which we employed signal detection theory to obtain empirical statistical measures of cognitive hacking conducted by mirage theory in a specific attack–defense model. In conclusion, this research overall showed that counter attack vectors that lie in defensive deception are a viable approach to protecting electrical power infrastructures from computer network attacks.

## REFERENCES

- [1] R.L. Krutz, *Securing SCADA Systems*, Wiley Publishing, 2006.
- [2] T. Roxey, Nuclear sector mitigation experience—an Aurora experience, in: *Process Control Systems Forum Annual Meeting*, August 2008 (Online). Available: [http://csrp.inl.gov/pcs/2008/d/nuclear\\_sector\\_mitigation-roxey.pdf](http://csrp.inl.gov/pcs/2008/d/nuclear_sector_mitigation-roxey.pdf).
- [3] US Joint Chiefs of Staff, *Joint doctrine for information operations*, Joint Publication 3-13, Defense Technical Information Center, October 1998.
- [4] M. Young, R. Stamp, *Trojan Horses—Deception Operations in the Second World War*, Bodley Head, London, UK, 1989.
- [5] N.C. Rowe, H. Rothstein, Deception for defense of information systems: analogies from conventional warfare. Technical Report of the Department of Computer Science and Defense Analysis, US Naval Postgraduate School, USA, 2003.
- [6] N.C. Rowe, H. Rothstein, Two taxonomies of deception for attacks on information systems, *Journal of Information Warfare* 3 (2) (2004) 27–39.
- [7] E. Montagu, *The Man Who Never Was*, Lippincott Publishing House, 1954.
- [8] L. Spitzner, *Honeypots: Tracking Hackers*, Addison-Wesley Professional, 2002.
- [9] HoneyNet Project, *Know Your Enemy: Learning About Security Threats*, 2nd ed., Addison-Wesley Professional, 2004.
- [10] T. Holz, J. Goebel, J. Hektor, *Advanced Honeypot-based Intrusion Detection*. *login*: 31 (6) USENIX, 2006.
- [11] C. Kreibich, J. Crowcroft, Honeycomb—creating intrusion detection signatures using honeypots, in: *Proceedings of the 2nd Workshop on Hot Topics in Networks*, Cambridge, MA USA, 2003.
- [12] N. Provos, *Developments of the Honeyd Virtual Honeypot* (Online). Available: <http://www.honeyd.org>.
- [13] T. Holz, F. Raynal, Detecting honeypots and other suspicious environments, in: *Proceedings of the 6th IEEE Information Assurance Workshop*, United States Military Academy, West Point, NY, USA, 2005.
- [14] J. Yuill, M. Zappe, D. Denning, F. Freer, *Honeyfiles: Deceptive files for intrusion detection*, in: *Proceedings of the 5th IEEE Workshop on Information Assurance*, US Military Academy, West Point, NY, USA, 2004.
- [15] N.C. Rowe, Finding logically consistent resource-deception plans for defense in cyberspace, in: *Proceedings of the 3rd International Symposium on Security in Networks and Distributed Systems*, Niagara Falls, Ontario, Canada, 2007.
- [16] US Joint Chiefs of Staff, *Military deception*, Joint Publication 3-13.4, Defense Technical Information Center, July 2006.
- [17] G. Giani, P. Thompson, Cognitive hacking: a battle for the mind, *IEEE Computer* 35 (8) (2002) 50–56.
- [18] C.S. Jones, The perception management process. *Military Review—The Professional Journal of the US Army*, 1999 (Online). Available: [http://www.au.af.mil/au/awc/awcgate/milreview/jones\\_perception.pdf](http://www.au.af.mil/au/awc/awcgate/milreview/jones_perception.pdf).
- [19] T.L. Thomas, Russia's reflexive control theory and the military, *The Journal of Slavic Military Studies* 17 (2004) 237–256.
- [20] F.E. Cellier, E. Kofman, *Continuous System Simulation*, Springer, 2006.
- [21] Mathworks Inc., *Simulink* (Online). Available: <http://www.mathworks.com/products/simulink/>.
- [22] Mathworks Inc., *Real-Time Workshop* (Online). Available: <http://www.mathworks.com/products/rtw/>.
- [23] D.F. Hoeschele, *Analog-to-Digital and Digital-to-Analog Conversion Techniques*, 2nd ed., Wiley-Interscience, 1994.
- [24] A.A. Cárdenas, S. Amin, S. Sastry, Research challenges for the security of control systems, in: *Proceedings of 3rd USENIX Workshop on Hot Topics in Security*, San Jose, CA, USA, July 2008 (Online). Available: [http://www.usenix.org/event/hotsec08/tech/full\\_papers/cardenas/cardenas.pdf](http://www.usenix.org/event/hotsec08/tech/full_papers/cardenas/cardenas.pdf).
- [25] J. Larsen, *Breakage*, Blackhat Federal, 2008.
- [26] J.L. Rrushi, K.D. Kang, *CyberRadar: A regression analysis approach to the identification of cyber-physical mappings in process control systems*, in: *Proceedings of the IEEE/ACM Workshop on Embedded Systems Security*, Atlanta, Georgia, USA, 2008.
- [27] S.M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*, Prentice Hall Publishing, 1998.
- [28] J.I. Marcum, A statistical theory of target detection by pulsed radar, US Air Force, Project RAND, December 1947 (Online). Available: [http://www.rand.org/pubs/research\\_memoranda/2006/RM754.pdf](http://www.rand.org/pubs/research_memoranda/2006/RM754.pdf).
- [29] WinTECH Software Design, *Modbus Master Data Scanner* (Online). Available: <http://www.win-tech.com/demos/modscan32.zip>.
- [30] Modbus Organization, *Modbus application protocol specification*. (Online). Available: [http://www.modbus.org/docs/Modbus\\_Application\\_Protocol\\_V1\\_1b.pdf](http://www.modbus.org/docs/Modbus_Application_Protocol_V1_1b.pdf).
- [31] Wireshark Foundation, *Wireshark* (Online). Available: <http://www.wireshark.org/>.
- [32] H. Stanislaw, N. Todorov, Calculation of signal detection theory measures, *Journal of Behavior Research Methods, Instruments, & Computers* 31 (1) (1999) 137–149.