Using Deception to Facilitate Intrusion Detection in Nuclear Power Plants
Julian L. Rrushi\textsuperscript{1,2} and Roy H. Campbell\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, University of Illinois at Urbana-Champaign, 201 N. Goodwin Avenue, Urbana, IL 61801, USA
\textsuperscript{2}Dipartimento di Informatica e Comunicazione, Università degli Studi di Milano, Via Comelico 39/41, I-20135, Milano, Italy
jrrushi@uiuc.edu
rhc@uiuc.edu

Abstract: In this paper we propose reactor mirage theory as a deception-based intrusion detection approach for digital I&C systems in nuclear power plants (NPPs). We draw from military deception techniques based on simulation of physical targets such as troops, radar-equipped air defense installations, tanks, bridges, airfields, etc. We propose the employment of genuine digital I&C systems to simulate physical components of a NPP via generation of Modbus protocol data units (PDUs) typical to the operation of these components. Communicating finite state machines are used to generate and recognize such deceptive PDUs. Artificially generated Modbus traffic is the reactor mirage theory counterpart of electromagnetic beam reflections, heat emitters, etc., commonly used as deceptive mechanisms by the military in warfare to indicate the existence of physical targets. These deceptive PDUs produce a drastic incrementation of the uncertainty which attackers may be subject to during the selection of target NPP components they plan to hit, hence increase by a high order of magnitude the probability of detection of attacks on NPP components.

Keywords: MILDEC, intrusion detection, digital I&C systems, nuclear power plants, signal detection theory, reactor mirage theory

1. Introduction
Cyber security exercises such as Cyber Storm (U.S. Department of Homeland Security, 2006) have demonstrated that it is indeed practically possible to intrude into process control networks (PCNs) monitored and operated via digital instrumentation and control (I&C) systems, i.e. microprocessor based systems, field programmable gate arrays, application specific integrated circuits, etc. Further, studies on the field report that process control systems in general may be subject to vulnerabilities in their data, security administration, architecture, networks and platforms (Stamp et al, 2003), and that cyber attacks on a NPP have the potential of causing serious physical damage (Krutz, 2006). Monitoring and control functions in NPPs were originally provided by analog devices, consequently cyber attacks were not applicable to them. Nevertheless, generation III+ and IV reactors are being built with digital I&C systems, and in older reactors analog devices are being replaced with digital I&C systems due to aging factors. In this paper we address the problem of intrusion detection in NPP control networks. The basis of our work, which we refer to as reactor mirage theory, is formed by deceptive capabilities derived from military deception techniques commonly used in warfare. Further, our work is developed under the guidance of Signal Detection Theory (Kay, 1998; Marcum, 1947). The contribution to intrusion detection made by the application of reactor mirage theory on a specific NPP is measured through the probability of detection, which in turn is defined as the probability that an attacker will send a control frame to a process interface device to query or change the status of a dummy physical component of a NPP. The proposed approach has shown to be promising due to a high degree of uncertainty injected into various steps of potential attacks.

The remainder of this paper is organized as follows. Section 2 provides an overview of relevant related work. Section 3 provides a detailed description of the proposed intrusion detection approach, namely reactor mirage theory. Section 4 provides a practical and model-based evaluation of the intrusion detection capabilities provided by reactor mirage theory. Section 5 summarizes our findings and concludes the paper. For the sake of simplicity in this paper we use the word attack to refer to cyber attacks on digital I&C systems in a NPP. In this paper we take Modbus (Modbus Organization, 2006) as an
example industrial protocol used by digital I&C systems in a NPP. Further, the NPP we refer to in this paper is a boiling water reactor.

2. Related work
The most relevant work related to reactor mirage theory is military deception (MILDEC) and its applications (Young and Stamp, 1989). 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 (U.S. Joint Chiefs of Staff, 2006). The deception means in MILDEC are grouped into three categories, namely physical means such as dummy and decoy equipment and devices, tactical actions, movement of forces, etc., technical means such as emission of chemical or biological odors, radiation, reflection of energy, etc., and administrative means such as techniques to convey or deny physical evidence. Reactor mirage theory is quite similar to MILDEC to a degree that could allow it to be thought of as almost an application of MILDEC to cyber security of cyber-physical systems such as NPPs.

Reactor mirage theory employs physical means of deception, namely genuine digital I&C systems which appear to be monitoring and controlling existing physical components of a NPP that in reality are dummy. Reactor mirage theory also employs technical means of deception, namely “emission” of artificially created control traffic, to deceive potential attackers into believing that behind a defined set of digital I&C systems there are a series of physical components of a NPP. One of the applications of MILDEC which best highlights the conceptual relation between MILDEC itself and reactor mirage theory is the Fortitude South operation which took place during world war II (Young and Stamp, 1989). During the allied invasion of German occupied territory of France, a special electronic unit called the 5th wireless group was created to deceive the command of German military into believing that the allies would attack through Pas de Calais rather than through Normandy.

The 5th wireless group used newly developed transmitters to generate radio communications according to some especially developed scripts. The radio traffic generated contained conversations typical to military assault operations. Under those conditions, the German military in France had no aerial reconnaissance capabilities therefore eavesdropping on allies radio communications was the principal mechanism they had to determine movements of allies troops. Reactor mirage theory exploits similar concepts, namely adversaries' reliance on analysis of intercepted traffic that is typical to the activity of those physical objects, and the impossibility of adversaries to verify that intercepted traffic is indeed generated by existing physical objects. Honeypots share a few features with reactor mirage theory to some extent. Honeypots are closely monitored information system resources serving as network decoys (Spitzner, 2002; The Honeynet Project, 2004).

Honeypots are employed for purposes such as distracting potential attackers from valuable machines, providing early warning about new attacks, and studying attackers and their techniques. Like honeypots, reactor mirage theory aims at diverting attackers' attention from valuable targets. Further, just like honeypots, the digital I&C systems used by reactor mirage theory may not have any production values. Nevertheless, unlike honeypots which are passive and just stand by to receiving connections, reactor mirage theory uses digital I&C systems which actively generate, transmit, and recognize network traffic. Unlike honeypots which simulate computer services that are often more vulnerable than their production counterpart in order to lure attackers, reactor mirage theory simulates physical components in such a way as to make these dummy components indistinguishable from existing components.

Unlike honeypots whose deception capabilities are placed within boundaries of the area in a computer system which is possibly visible to attackers through network access, reactor mirage theory places simulation capabilities at a layer which is not reachable by potential attackers via cyber access to a target control network.

3. Reactor mirage theory
We now provide a description of intrusion detection capabilities for digital I&C systems based on deceptive factors and developed under the light of SDT. Although a good part of
the work on SDT was carried out in radar research, SDT has a direct application to our work with regard to analysis of the decision making of attackers during target selection. The ultimate goal of the implementation of attacks on a NPP is to cause physical damage to critical NPP components such as reactor core, electric power generators, etc. Technically a successful implementation of attacks on a NPP requires the identification of a defined digital I&C system controlling a target NPP component and knowledge of the operation of this digital I&C system on the target NPP component. For example, an attack carried out to cause a loss of reactor coolant requires the identification of one or more digital I&C systems which implement control functions on reactor feed pumps. These feed pumps raise the pressure of the moderator, i.e. water, in order to make it flow to the reactor vessel.

Furthermore, such an attack requires identification of what Modbus data items in the memory of a digital I&C system represent feed pump operational conditions such as rates and status. Attackers are also required to find out the number of feed pumps in operation in a target NPP. Blocking the operation of only a part of these feed pumps, while the rest of them are in operation, may not cause a sufficient drop in reactor water level to initiate physical damage to the reactor core. An underlying principle of SDT is that decision making takes place under a defined quantity of uncertainty. In the case of an attack targeting loss of reactor coolant, for example, the decision that a defined I&C system with IP address 10.0.1.24 controls a defined feed pump, that a holding register with Modbus address 54 in this system is associated with the rate of this feed pump, and that the contribution to the reactor water level provided by this feed pump is 0.25%, is taken by an attacker under some uncertainty.

Attackers acquire information on target NPP components through analyses of control traffic flowing over PCNs. Examination of Modbus PDUs, for example, reveals the content of Modbus data items stored in the memory of a digital I&C system and Modbus addresses of these data items. By comparing these values to typical values of operational variables of equipment such as neutron monitoring sensors, temperature sensors, pressure sensors, instrument transformers, etc., and valves, pumps, circuit breakers, motors, etc., used by digital I&C systems to monitor and operate on a NPP, respectively, attackers can derive the aforementioned information necessary for a target selection. PDUs with function codes 0x04 and 0x06, and register values 0x384 and 0x54, for example, are likely to denote that the reactor pressure is about 900 pounds per square inch, and operators are setting the rate of a feed pump to its maximum capacity, namely 8.4 units of millions of pounds per hour (MLB/hr), respectively.

![Figure 1: POC curves for a digital I&C system operating on a NPP via reactor feed pumps estimated during a red team attack on a simulated NPP targeting loss of reactor coolant](image)
Binding the semantics of these PDUs to the IP addresses of the digital I&C systems which have sent and have received the PDUs in question, respectively, reveals the functionalities that are likely to be carried out by these digital I&C systems in the overall operation of a NPP. Although from the target identification point of view there is some of what in SDT is referred to as external noise, such as for example various overlaps between typical values of different equipment used to monitor and control NPP components, the uncertainty under which an attacker decides that such derivation is correct and usable is relatively low. Figure 1 depicts the internal response probability of occurrence (POC) curves characterizing the uncertainty under which a set of red team individuals identify a defined digital I&C systems as target of an attack whose objective is the causation of a loss of reactor coolant in a simulated NPP. The horizontal axis represents information which motivates red team members to decide that the signal is present, while the vertical axis represents the frequency of the occurrence of a defined amount of this information.

The POC curves in Figure 1 show that signal strength is high and the amount of noise is low. Consequently the overlap of these curves is limited and their spread is reduced, leading to a discriminability index of $d' = 5.6$. These estimation results show that attackers who possess expertise of digital I&C systems, industrial protocols used in NPPs, equipment used to operate on NPP components, and NPP operation, can identify their targets with a high rate of correct selections and a low rate of wrong selections. Figure 2 depicts the receiver operating characteristic (ROC) curve describing such rates. The approach proposed in this paper produces a drastic incrementation of the uncertainty which attackers may be subject to during the selection of target NPP components they plan to hit, in order to increase the probability of detection of attacks on NPP components.

![Discriminability index $d' = 5.6$](image)

**Figure 2:** ROC curve corresponding to the POC curves of Figure 1

The basis of the proposed approach is formed by a simulation of the existence of a large set of physical components of a NPP. We draw from military deception techniques based on electromagnetic beam reflections, heat emitters, plastic and wooden objects, bonfires, etc., to simulate the existence of physical targets such as troops, radar-equipped air defense installations, tanks, bridges, airfields, etc. From the perspective of cyber access to a PCN the existence of each physical component of a NPP actually in effective operation reflects Modbus traffic sent to, or generated by, digital I&C systems acting as process interface devices in a NPP. As written earlier in this paper, attackers analyze such control traffic in order to identify potential targets of their attacks. Transmissions over a PCN of Modbus request PDUs writing a set of pressure regulating Modbus data items...
stored in a defined digital I&C system to control main turbine control valves or turbine bypass valves, for instance, indicate the existence of a steamline. Similarly, Modbus response PDUs carrying status data generated by neutron monitoring sensors indicate the existence of nuclear fuel, i.e. a set of Uranium rods, and fission control rods. Transmission over a PCN of Modbus response PDUs carrying current and voltage measurement data measured by instrument transformers and acquired via intelligent electronic devices (IEDs) indicate the existence of electric power generators. We propose the employment of genuine digital I&C systems to simulate physical components of a NPP via generation of Modbus request PDUs and Modbus response PDUs typical to the operation of these physical components. Artificially generated Modbus traffic is the reactor mirage theory counterpart of electromagnetic beam reflections, heat emitters, etc., commonly used as deceptive mechanisms by the military in warfare to indicate the existence of physical targets.

The presence of a large set of artificially generated deceptive PDUs in Modbus traffic examined by attackers during target identification deceives them into deriving not only the existence of physical components of a NPP, which in reality do not exist, but also information on digital I&C systems that appear as being used to monitor and control these components, configuration of the Modbus protocol in these digital I&C systems, the semantics of Modbus data items stored in the memory of these digital I&C systems, and the equipment, i.e. pumps, valves, etc., which appear as being used by these digital I&C systems to operate on a NPP. Consequently there is a probability that attackers select and subsequently attack dummy NPP components. Larger the set of simulated NPP components, higher the probability that attackers will end up with attacking dummy NPP components. The proposed approach may be even extended to simulate entire NPPs. Deceptive PDUs have no effects on the hit rate. If signal is present, i.e. an existing physical component of a NPP is indeed the target being sought by attackers, and attackers select this NPP component as a target of their attacks, deceptive PDUs will not affect their decision to do so. Nevertheless, deceptive PDUs drastically increase the false alarms rate. If the signal is absent, i.e. what may appear as the target being sought by attackers is not what attackers are looking for, deceptive PDUs will affect attackers' decision to select this sort of mirage as the target of their attacks. We introduce deceptive PDUs as much as possible in the form of internal noise. Attackers can do little or nothing to reduce such noise as the proposed approach exploits the fact that the physical part of a digital I&C system acting as a process interface device is not visible to cyber access over a PCN.

![Diagram of NPP simulation](image)

**Figure 3:** NPP simulation used to test CFSMs developed through the CHSM language system
The cyber part and the physical part in digital I&C systems interact closely with each other, but still are different in nature. A remote access to digital I&C systems over a PCN may provide visibility on the cyber part of this digital I&C system. This visibility covers also measurement digital data generated by the physical part and forwarded to the cyber part, and control data forwarded to the physical part from the cyber part. Nevertheless, the access in question provides no insight into the physical part of a digital I&C system. There is no way through cyber access to determine whether data generated from the physical part of a digital I&C system are the result of the fact that this system is attached to a physical component of a NPP and is operating on it, or these data are artificially generated in order to mimic the operation of this digital I&C system on a physical component of a NPP.

The rate of neutron production in the fuel region, for example, is provided by digital I&C systems upon reception of signals transmitted by detectors located within the reactor core. The sensing capabilities of these sensors are based on physical processes. Cyber access to the digital I&C systems in question has no means of providing any insight into detectors within the reactor core. Most importantly, cyber access does not allow for verifying that a series of physical processes based on interactions with neutrons indeed took place in a detector within an existing reactor core. The same holds for actuators, i.e. pneumatic, hydraulic, or electrically powered devices that provide motion of equipment such as valves or pumps. Cyber access to digital I&C systems used to control actuators has no means of providing for a verification of the existence of valves or pumps along with other related physical components of a NPP.

**Figure 4:** Excerpt from two CFSMs generating and recognizing Modbus traffic related to a controlled change of the rate of a reactor feed pump
We exploit the possibility that under such massive internal noise attackers may identify dummy NPP components as targets of their attacks and subsequently attack them, to build an intrusion detection mechanism. The proposed approach employs communicating finite state machines (CFSM) (Heimdahl, 1998; Romdhani et al 1995; Tropper and Boukerche, 1993) to deterministically and consistently generate valid Modbus traffic appearing as being part of the operation of existing physical components of a NPP. In addition to traffic generation, these CFSMs are also employed to recognize artificially generated traffic. Thus, a state machine $M_1$, which is a member of the CFSMs in question, generates Modbus traffic that is to be recognized by one or more state machines, say $M_2$ and $M_3$, which are members of these CFSMs as well. On the other hand, state machines $M_2$ and $M_3$ generate Modbus traffic that is to be recognized by state machine $M_1$.

We implemented CFSMs through the Concurrent Hierarchical State Machine (CHSM) language system (Lucas and Riccardi) and tested their ability to correctly generate and recognize Modbus traffic in the simulated environment depicted in Figure 3. The simulated digital I&C systems in Figure 3 are based on VMWare and run the uCLinux operating system (Albanowski and Dionne) along with a free version of an implementation of the Modbus protocol, namely FreeModbus (Walter). For the sake of simplicity of testing, PCNs and Fieldbuses are merged into a single simulated network. Various sensors and actuators are also simulated via FreeModbus. In Figure 4 is given a sample generation and recognition of Modbus traffic. Figure 4 depicts an excerpt from CFSM $M_1$ used in a Modbus master device, and CFSM $M_2$ used in a Modbus slave device attached to a reactor feed pump.

Upon reception of a starting message, $M_1$ generates a Modbus PDU with function code 0x06, i.e. write single register, register address 0x4e85, and register value 0x3c, and transitions from initial state $S_{10}$ to state $S_{11}$. CFSM $M_2$ receives in input the output of CFSM $M_1$, generates a Modbus PDU identical to the one received from CFSM $M_1$, and transitions from initial state $S_{20}$ to state $S_{21}$. Note that input and output of CFSM $M_2$ are identical since according to the specification of Modbus protocol a Modbus slave device generates a Modbus response PDU identical to the Modbus request PDU previously received, to indicate a successful single write of a holding register (Modbus Organization, 2006). CFSM $M_1$ receives in input the output of CFSM $M_2$, generates a Modbus PDU with function code 0x06, register address 0x4e85, and register value 0x50, and transitions from state $S_{11}$ to state $S_{12}$. CFSM $M_2$ receives in input the output of CFSM $M_1$, generates...
a Modbus PDU identical to the one received from CFSM M1, and transitions from state S21 to state S22.

CFSM M1 receives in input the output of CFSM M2, generates a Modbus PDU with function code 0x06, register address 0x4e85, and register value 0x00, and transitions from state S12 to state S10. CFSM M2 receives in input the output of CFSM M1, generates a Modbus PDU identical to the one received from CFSM M1, and transitions from state S22 to state S20. Attacks on dummy NPP components result in transmission of Modbus PDUs which modify Modbus data items stored in the memory of a digital I&C system with values that initiate physical damage to a target NPP. With this regard, in addition to generation of Modbus traffic, the simulation of physical components of a NPP also includes consistent responses to offensive PDUs, hence deceiving attackers into thinking that their attack is taking the target NPP to abnormal operational conditions.

If attackers target a loss of reactor coolant, for example, and send offensive PDUs to a digital I&C system controlling a set of feed pumps which in reality are only simulated and do not exist, a digital I&C system monitoring a dummy reactor core will respond with deceptive Modbus response PDUs reporting a rapid drop of the reactor water level. As CFSMs themselves generate Modbus PDUs characterizing a normal operation of a NPP, these CFSMs do not recognize PDUs injected by attackers. Furthermore, these CFSMs do not recognize PDUs which could have been part of a normal operation of a NPP but do not belong to the ordered set of PDUs preliminarily scheduled to simulate physical components of a NPP. Deceptive capabilities provided by the approach proposed in this paper include simulation of the occurrence of NPP transients at a proper rate. Transients are abnormal events such as loss of feed water heating, inadvertent initiation of a coolant injection system, pipe breaks, electrical faults, etc., which perturb important NPP operational variables. In most cases a NPP can be taken to normal conditions without having to initiate a scram, i.e. a rapid shutdown of a NPP. The failure rate of physical components of a NPP may be used by attackers to discern between existing and simulated NPP components. Therefore it is highly critical to simulate the occurrence of transients at a consistent rate. We employ techniques such as condition data monitoring and aggregation provided in (McCalley et al, 2006; Pathak et al, 2005; Zhang et al, 2006) to estimate the failure rate of electric power generators and most of the other physical components of a NPP. The overall deceptive capabilities provided by the proposed approach generate an internal noise, which, as shown in the following section, increases the probability of detection to a high order of magnitude.

Figure 6: ROC curve corresponding to the POC curves of Figure 5
4. Evaluation

The proposed intrusion detection approach comes with a financial cost since it requires the deployment of additional genuine digital I&C systems, possibly of the same kind and configuration as the digital I&C systems used to monitor and control existing physical components of the NPP to be protected. Nevertheless, replication is a common practice in NPPs, therefore the requirement of additional digital I&C systems posed by the proposed approach may not sound unusual to the nuclear power industry. As simulation of physical components of a NPP is carried out through generation of Modbus traffic, the proposed approach regularly introduces network overhead. Nevertheless, this cost may be quite affordable taking into account the deployment of high speed and high bandwidth networking technology in NPP PCNs.

The internal noise introduced by the proposed approach causes a drastic increase of the uncertainty under which attackers identify the targets of their attacks. Such a high uncertainty provides a high probability that attackers will end up with attacking dummy NPP components and consequently get detected by CFSMs. Figure 5 depicts the internal response POC curves characterizing the uncertainty under which the aforementioned red team identifies a defined digital I&C systems as target of an attack targeting a loss of reactor coolant. POC curves of Figure 5 show that in a NPP equipped with deceptive capabilities provided by the proposed approach the signal strength is low and the amount of noise is high. After applying the proposed approach the discriminability index representing a measure of the strength of the internal response is lowered from $d' = 5.6$ to $d' = 0.45$. Hit rates and false alarm rates of a red team during target selection under the effect of deceptive capabilities are provided by the ROC curve of Figure 6.

Referring to Figure 7, let $\theta$, $\sigma_i$, and $\tau_i$ denote the probability of detection, the probability of selection of $PCNi$ out of, say, 4 PCNs as a network leading to target digital I&C systems, and the probability of selection of a dummy NPP component in $PCNi$ as the target of an attack, respectively. Let $x_i$ and $y_i$ denote the number of dummy NPP

![Figure 7: Example of a NPP equipped with deceptive capabilities provided by the proposed approach](image_url)
components and the number of existing NPP components monitored and controlled over \( PCNi \), respectively. The probability of detection \( \theta \) is given by:

\[
\theta = (\alpha_1 \times \tau_1) + (\alpha_2 \times \tau_2) + (\alpha_3 \times \tau_3) + (\alpha_4 \times \tau_4)
\]

If \( x_1 = 160, y_1 = 240, x_2 = 400, y_2 = 0, x_3 = 400, y_3 = 0, x_4 = 400, y_4 = 0 \), a discriminability index so close to zero means that the probability that attackers get detected only because they select a dummy NPP component as the target of their attack is slightly lower than 0.85. The efficiency of the proposed approach may be further confirmed by estimating the reward accumulated over an interval of time. We employ the stochastic activity network (SAN) formalism (Sanders and Meyer, 2001) through the Möbius tool (Deavours et al., 2002; Sanders, 1999) to build a model of attackers' advances in a NPP until selecting a NPP component as a target of their attacks. The SAN model in question is depicted in Figure 8.

Attacks along with PCN selections and NPP component selections are modeled as timed activities, while attacker decisions are modeled as SAN places. In Möbius the rates of timed activities are parameterized via global variables. The SAN model in question is then solved each time a new global variable assignment is applied. Such a Möbius study (Deavours et al., 2002) allows for estimating the rewards accumulated upon application of various activity time distribution functions for the occurrence of attacks, PCN selections, and NPP component selections. We define rates of reward that belong to the Interval-of-Time category of the activity-marking reward structure (Sanders and Meyer, 1991) in the SAN model depicted in Figure 8. These rates of reward are associated with tokens in SAN places modeling selections of dummy NPP components and are estimated through solutions of the SAN model in question.

![Figure 8: A SAN model of attack target selections](image)

### 5. Conclusion

In this paper we propose reactor mirage theory, i.e. an intrusion detection approach whose operation is based on the uncertainty which attackers are subject to during the identification of physical components of a NPP as targets of their attacks. We demonstrate the feasibility of simulating physical NPP components in a cyber way, and highlight the benefits of such deceptive capabilities from the intrusion detection point of view. We show that reactor mirage theory has the potential of reducing discriminability index, i.e. a measure of the strength of the internal response, from \( d' = 5.6 \) to \( d' = 0.45 \). We also provide an estimation of the intrusion detection probability provided by reactor mirage theory in an example NPP. In conclusion, we use the SAN formalism to model attack target selections in order to gain further insight into the efficiency of the proposed approach by estimating the reward accumulated over an interval of time.
5. Acknowledgement
Julian Rrushi was supported in part on a doctoral scholarship from the Università degli Studi di Milano, and in part on a research scholarship from (ISC)². Opinions, findings and conclusions expressed in this paper are those of the authors only.

References


Zhang, Z., Jiang, Y. and McCalley, J.D. (2006) *Bayesian Analysis of Power Transformer Failure Rate Based on Condition Monitoring Information*, Iowa State University, research sponsored by Power Systems Engineering Research Center, USA.