The Application of Virtual Proofs of Reality to Nuclear Safeguards and Arms Control Verification

Sébastien Philippe,[*,1] Moritz Kütt,[*] Michael McKeown,[*]
Ulrich Rührmair,[†] and Alexander Glaser[*]

[*] *Princeton University, Princeton, NJ, USA*

[†] *Horst Görtz Institute for IT Security, Ruhr-University Bochum, Germany*

[1] Corresponding author: sp6@princeton.edu

ABSTRACT. Nuclear inspections, for example as part of non-proliferation treaty safeguards and arms control agreements, often involve compliance verification in sites to which inspectors have limited access. Traditionally, this necessitates the irreversible transfer of inspector-provided sensor equipment to the inspected parties prior to inspection. This can lead to competing interests between inspectors and inspected parties, for example concerning the frequency or intrusiveness of onsite inspections, and regarding the question of mutual trust in the sensor equipment. Meeting these requirements continues to be a challenge. This paper develops a radically new approach to this problem based on the idea of "Virtual Proofs of Reality." Virtual Proofs offer a way to prove physical statements over insecure communication channels between two parties in two separate locations. They do not require classical tamper-resistant sensor hardware with cryptographic keys to this end, but rely on the use of Physical Unclonable Functions (PUFs) in an interactive protocol instead. Among other things, this reduces the necessary security assumptions on the sensor equipment, making secure sensor fabrication and mutual trust in the equipment substantially easier. Our paper discusses two example Virtual Proofs that appear particularly relevant in the context of nuclear safeguards and arms control.

Introduction

Onsite inspections are a key mechanism for nuclear verification. They play a critical role in monitoring treaty compliance as well as deterring non-compliance. They are a very intrusive mechanism, however, and can be expansive and logistically complex to implement. This leads to competing interests between inspectors and inspected parties during negotiations, for example concerning the frequency and intrusiveness of inspections, which can eventually prevent or limit their implementation.[1] Furthermore, inspections that involve physical measurements in sensitive locations or on military sites require trusted sensor equipment. This can necessitate the irreversible transfer of such equipment from the inspecting party to the inspected party prior to inspection.[2] To reduce the frequency of inspections, both parties can agree on using unattended or

remote-monitoring sensors.[3] In both cases, sensor hardware and communication channels are assumed to be tamper resistant and trusted by the inspector to ensure that data acquired are trustworthy.[4] Demonstrating these underlying security assumptions continues to be a challenge.

Here, we propose a radically new approach to this problem that does not require classical tamper-resistant sensor hardware and cryptographic keys.[5] We rely instead on the use of Physical Unclonable Functions (PUFs) in interactive protocols where the host is the prover and the inspector the verifier. Such security protocols, recently introduced by Rührmair et al.,[6] are called "Virtual Proofs of Reality" (VPs). Virtual proofs are a class of interactive proof systems that allow proving physical statements over insecure communication channels between two parties in two separate locations without using classical secret keys and tamper-resistant hardware. In the context of nuclear safeguards and arms-control verification, they could provide a basis for conducting challenge inspections *from a distance*, especially for in places where inspector access is temporarily or permanently limited or impossible.

After introducing the general concept of Virtual Proofs, we present two concrete applications: First, as an example for a measurement of a physical observable, we discuss a virtual proof of temperature (measured at a remote location) using a Bi-stable Ring PUF (BR-PUF) based on a field programmable gate array (FPGA) following the work of Rührmair et al. (2015). Second, we introduce a new type of virtual proof that would demonstrate the viability of the concept for nuclear radiation measurements. Specifically, as a proof-of-concept, we propose a virtual proof of neutron non-irradiation (or absence of neutron irradiation) using non-electronic superheated emulsion detectors. As we discuss these examples, we highlight the potential for practical applications of virtual proofs in nuclear verification.

The Concept of Virtual Proofs[7]

In a VP, the prover and the verifier are located in two distinct physical systems S_1 and S_2 respectively and can communicate over an abstract and ideal digital channel. The prover in S_1 wants to prove a physical statement to the verifier in S_2. Examples of physical statements are the temperature or location of a physical object in S_1.

Like any other interactive proof system, a VP must be sound and complete. The soundness and completeness properties guarantee that the verifier cannot be tricked into accepting false statements (soundness) and that the prover will be able to convince the verifier to accept true statements (completeness).[8] Some additional assumptions about the physical systems can be made as well. For example, one can assume S_1 to be "closed" such that the system has no possibilities of external physical exchanges other than the communication channel.

Typically, a VP consists of two phases: a setup phase and a proof phase. The setup phase can be either public or private with regard to the prover. Here, we restrict our example to VPs with a private setup phase. In such a phase, the verifier prepares so-called "witness objects" and used them to measure relevant physical properties, storing privately the data obtained.

A particularly interesting class of witness objects for VPs are objects that behave as physical unclonable functions (PUFs) and have output dependent upon some physical quantities. A PUF is a *partly* disordered physical system that can be challenged with external stimuli, or challenges C_i, upon which it reacts producing responses R_i. Contrary to standard digital systems, a PUF's responses shall depend on the nanoscale structural disorder present in it. It is assumed that this disorder cannot be cloned or reproduced exactly, not even by the PUF's original manufacturer, and that it is unique to each PUF. This means that any PUF implements an individual function F_{PUF} mapping challenges C_i to responses R_i. The tuples (C_i, R_i) are called the challenge-response pairs (CRPs) of the PUF. With such properties, PUFs are considered to be the physical equivalent of one-way functions.[9] They are easy to evaluate but hard to predict, easy to manufacture but almost impossible to duplicate.

During the setup phase of a PUF-based VP, the verifier prepares a list $\mathcal{L} = (R_j^i, C_j^i, \Theta_j)$ of CRPs such that for a given value of a physical quantity Θ_j and a challenge C_j^i the PUF response R_j^i is given by $R_j^i = F_{PUF}(C_j^i, \Theta_j)$. Then, the witness objects are transferred to the prover's system S_1 to be used in the proof. The objects do not contain any cryptographic keys nor need to be assumed tamper-resistant in the classical sense. The prover can open and inspect the objects. If the prover uses destructive measures to do so, he will need to either physically clone or computationally simulate the witness object's physical behaviors before the proof takes place. Both actions are hard to implement when using PUFs. The prover can also attempt to evaluate $F_{PUF}(C, \theta)$ for all C and θ. However, the space of possible CRPs can be made arbitrarily large using adequate PUFs, preventing such an attack.

In the proof phase, the verifier and prover interact sequentially over the communication channel using a challenge-response protocol to demonstrate the state of the witness object in S_1. In a PUF-based VP, the prover would first send the value of the physical quantity to be proven, for example θ, to the verifier. The verifier would then look up his CRP list $\mathcal{L}$, identify $\Theta_j = \theta$, randomly pick i and send the challenge $c = C_j^i$ back to the prover. After receiving c, the prover sends back $r = F_{PUF}(c, \Theta_j)$. Finally, if $r = R_j^i$ the verifier accepts the proof. The verifier removes the (C_j^i, R_j^i) pair from $\mathcal{L}$.

In what follows, we turn our attention to two particular examples: first a VP of temperature independently reproducing experimental results from Rührmair et al. (2015), then a virtual proof of neutron irradiation and its potential implementation that we propose for the first time.

Virtual Proof of Temperature

In a VP of temperature, the prover wants to convince the verifier of the temperature of a witness object in S_1. Additional assumptions are needed for the proof to extend to the temperature of the system itself or of any subsystem or objects in S_1. This is an example of how physical sensors can be used in a VP.

Concept. The proof follows the general interactive protocol for PUF-based VP presented in the previous section. The PUF is assumed to be temperature dependent such that in the setup phase a challenge-response list $\mathcal{L}$ can be constructed as above for a set of discrete temperatures Θ_j. In the proof phase, the prover first claims that the PUF is at temperature $T \in \Theta_1, ..., \Theta_k$. Then, the proof continues at outlined in the previous section. If the proof is repeated r times, the dimension of $\mathcal{L}$ should at least be of size $\mathcal{O}(kr)$.

Apparatus. We test the underlying physical assumptions of the proof by implementing a BR-PUF[10] on a Xilinx Artix-7 35T FPGA. BR-PUFs are based on inverter rings consisting of inverter stages connected to one another. Each stage consists of two inverters and a pair of multiplexer and demultiplexer to select either of the inverters to be connected in the inverter loop. Challenges consist in choosing which inverter should be activated from each pair. Once the ring configuration is fixed, all connections are pulled low and released from this unstable state. It then falls into one of its two possible stable states. For a ring of four inverters, these states are "0101" and "1010." The response of the PUF is not only challenge dependent but also temperature dependent providing a basis for this VP. We implemented four 64-bit BR-PUFs and XOR their responses in the FPGA. We use four 16 pins MCP23O17 GPIO integrated circuits (ICs) connected by an I2C bus to deliver the 64 bits challenges. A fifth GPIO IC controls the FPGA and reads the response. The ICs are controled by an Adafruit FT232H Breakout board, allowing a PC to communicate to the I2C bus via usb connection. The complete apparatus is shown in figure 1. Python scripts are used to carry out the experiments.[11]

Experiment and results. We placed our FPGA board and a temperature sensor inside a cooler. Once the initial temperature stabilized to $\sim 12.5\,^{\circ}\mathrm{C}$, we ran a list of 200 64-bits challenges repeatedly 2000 times and averaged *a posteriori* the results for every 11 trials. The process of data collection lasted for about 20 hours, enough for the cooler to slowly equilibrate with the room temperature of $\sim 25\,^{\circ}\mathrm{C}$. The results of the challenges are presented in Figure 2. They show interesting behavior for the setup. First, not all challenges yield stable results. These are represented by green dots in the top left plot of figure 2. Second, most challenges are relatively uncorrelated and independent of temperature. However the remaining show enough variation to distinguish various discrete temperatures. For each temperature, it is possible to concatenate all

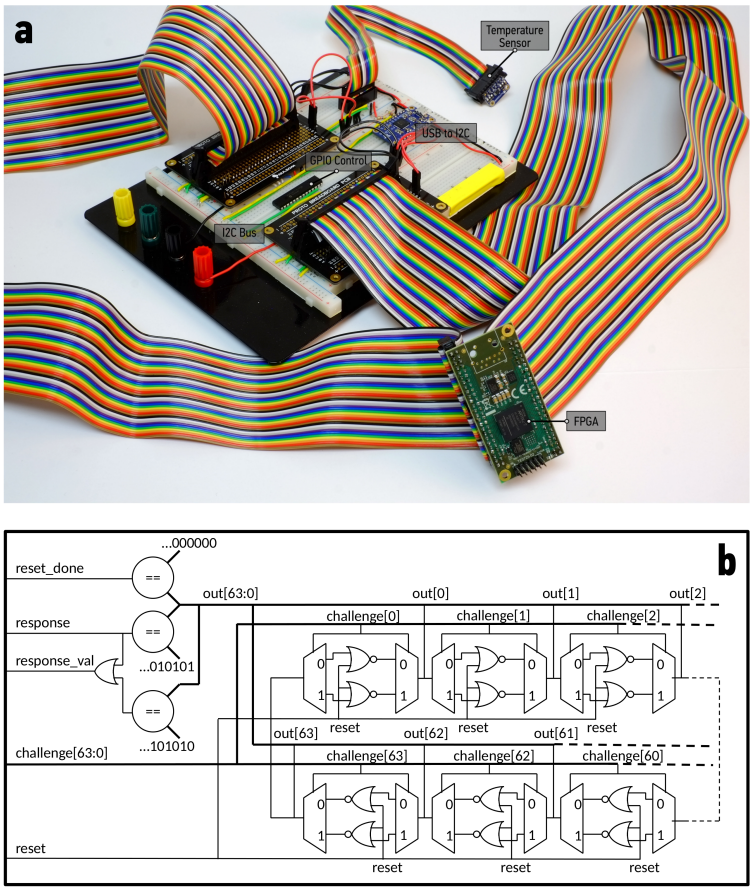

Figure 1. Virtual proof of temperature experiments. a) Hardware set-up comprising a FPGA, GPIO commands, I2C bus, USB to I2C and a temperature sensor. b) Schematic of the bi-stable ring physical unclonable function used in the experiment and coded on the FPGA.

challenges response into a 200-bit string. We can then calculate the hamming distance (HD) between the strings of any two temperatures. In figure 2, we show the HDs between six selected temperatures. We also show the intrinsic instability of our system that we define as the HD between two neighboring challenge strings. Results show that the instability is always smaller than the hamming distance between two of the selected temperatures. These results are consistent with what was previously reported by Rührmair et al. (2015).

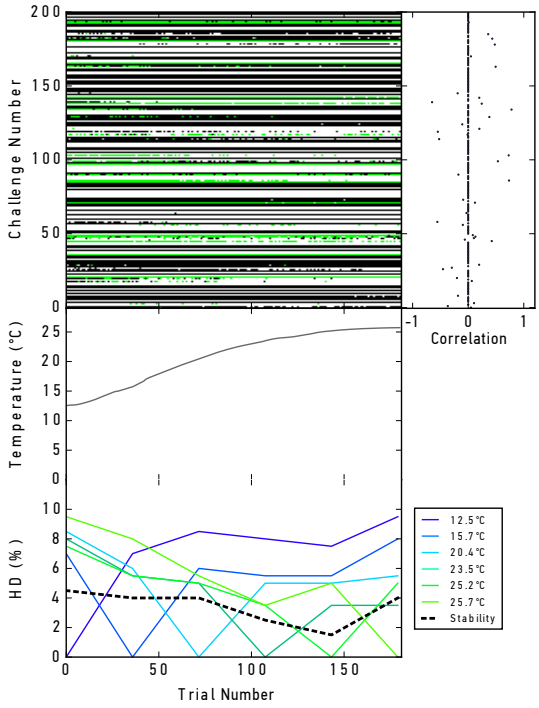

Figure 2. Results of the VP of temperature experiments. Every trial consists of 11 separate runs of 200 challenges. The top left panel shows the outcome of all trials (black = "1", white = "0" and green "unstable"). The top right panel shows the correlation between each individual challenges behavior and temperature change. The lower panel shows the hamming distances between the combined results of 200 challenges at six selected temperatures as well as the instability of our system defined as the hamming distance between two neighboring trials.

Virtual Proof of Neutron Non-Irradiation

The main objective of this project is to extend the concept of virtual proofs to the realm of nuclear verification and, in particular, to show that it can be applied to measurements of nuclear radiation, which are often central for inspections. Below, we propose a virtual proof of neutron non-irradiation; this particular proof strictly serves as a proof-of-concept and is not immediately meant to support or enable a particular verification scenario.

In a VP of neutron non-irradiation, the prover wishes to demonstrate to the verifier that a witness object has not been exposed to neutron radiation above a certain neutron energy threshold. Additional assumptions, for example, about the witness object location in S_1 could be used to demonstrate that neutron sources have not been removed

from a room or that no neutron sources have been introduced in a room—giving a basis for the application of VPs to perimeter monitoring in arms control and disarmament verification.

Concept. Choosing an appropriate witness object is crucial for this application. Here, we propose to use superheated droplet neutron detectors as a non-integrated optical PUFs (Figure 3). Optical PUFs are physically disordered media, typically a transparent object doped with scattering particles, which will produce a random and unique speckle pattern when exposed to a coherent light source.[12] Their security lies on the complex light-scatterers interactions inside the PUF. Speckle pattern outputs depend on the size and density of scatterers, the wavelength and diameter of the laser beam, and the orientation and position of the PUF with respect to the laser beam. So far, even sophisticated attacks based on machine learning algorithms have been unable to predict the outputs of non-integrated optical PUF.[13] Optical PUFs developed by Sandia National Laboratories have been used as a way to uniquely identify tags and seals used in safeguards and arms control applications.[14].

Superheated droplet neutron (or bubble) detectors have been used recently in the experimental demonstration of a physical zero-knowledge interactive proof for warhead confirmation.[15] They comprise standard glass vials filled with an emulsion of fluorocarbon superheated droplet suspended in an aqueous gel matrix.[16] Typical detectors have about 4000 droplets of 80–120 μm diameter per cubic centimeter and a volume of 8 cm^3. The detectors are insensitive to incident neutrons with energies below a certain threshold depending on the droplet and gel composition. They are insensitive to gamma radiation. When a metastable droplet vaporizes and explodes due to an energetic enough neutron interaction, it expands into a stable bubble about six times larger in diameter ($\sim 600\,\mu$m). The detectors can be irradiated several times and record the total fluence to which they are exposed.

Here we assume that the irradiation of a bubble detector with neutrons will modify the internal structure of the detector significantly and affect the speckle pattern. As soon as macroscopic bubbles are formed, the scattering property of the detector should change because bubbles are much larger and in a different phase than droplets. Moreover, by expanding, bubbles modify locally the distribution and locations of droplets.

Bubbles can be re-condensed by placing a detector in an isostatic pressure chamber at 500 psi for $\sim$ 10 minutes. While recompression will remove the macroscopic bubbles, it may or may not "re-initialize" the location of all droplets to their original positions (before irradiation). We address this issue by irradiating the detector to preload them with a small number of bubbles. If the prover tries to cheat and compress the detector, he would need to re-create the same bubble configuration. Particle interaction being a random process, the probability that for example the same five droplets out of 32,000 will expand again into the same bubbles is $\sim\ 1\,/\,2.79\ \times\ 10^{20}$. Bubble detectors are

sensitive to cosmic neutrons, an effect that could affect the measurements if detectors are to be used for long periods of time. Experiments with our most sensitive detectors showed that they accumulate on average $\sim \frac{2}{3}$ bubbles per detector per day at room temperature.

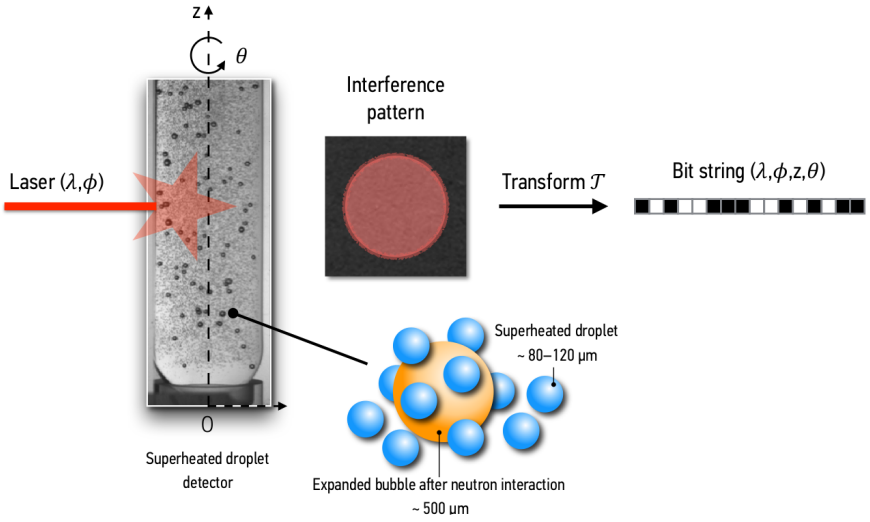

Figure 3. Concept of the virtual proof of neutron irradiation based on Pappu's physical one-way function. The superheated droplet detector behaves as an optical puf sensitive to neutrons. A laser with wavelength λ and beam diameter ϕ) probes the detector at various height z and angle θ. The resulting interference pattern is recorded with a CCD camera. A transform is applied to this image to obtain a challenge response (bit string).

Protocol for VP of Neutron Non-Irradiation.

Setup phase:

- The verifier prepares a bubble detector based optical PUF, and preload it with some bubble.
- She determines a private CRP-list $\mathcal{L}$ for the detector. For $i = 1, ..., n$, she randomly chooses challenges $C_i = (q_i, \theta_i)$, directs a laser beam at coordinate q_i with angle θ_i, and measures the resulting optical responses R_i behind the PUF.
- She transfers the detector to the prover.

Proof phase:

- The prover claims that the detector has not further been exposed to neutrons.
- For $v = 1, ..., m$, with $m < n$ the verifier randomly selects (C_v, R_v) pairs and send $C_v = (q_v, \theta_v)$ to the prover. For each C_v the prover directs the laser beam to the PUF according to (q_v, θ_v) and send the resulting optical response R_v^* back to verifier.
- If all $R_v^* = R_v$, the verifier accept the proof. She then removes the (C_v, R_v) pairs from the list $\mathcal{L}$.

Discussion. In theory, every unit volume of the size of the laser wavelength can have an influence on the optical response such that a system using a 4 cm long and 1.6 cm diameter detector (standard vial) exposed to a 600 nm laser has $\mathcal{O}(10^{13})$ available challenges. A VP of neutron non-irradiation could have many interesting applications such as perimeter monitoring for warheads dismantlement, naval fuel safeguards,[17], plutonium objects or other neutron sources. It could also be used in combination with other VPs such as the VP of distance based similarly on optical PUFs (Rührmair et al., 2015).

Conclusion and Future Work

Onsite inspections currently are a central element of nuclear safeguards and arms-control verification. They remain important and are probably irreplaceable, also as a confidence building measure for the parties to a treaty. There are many situations, however, where onsite inspections are difficult or perhaps even impossible to implement due to their intrusiveness and related security and safety concerns that a host party might have. This aspect is particularly relevant in the context of nuclear arms-control verification, especially for future treaties that might have to account for individual nuclear warheads. In the process of inspecting them, highly classified information could be placed at risk. The specter of highly intrusive onsite inspections could ultimately even prevent further progress in this area should negotiating parties conclude that the required tradeoffs (between treaty objectives and their verifiability) are not in their interest.

To address these concerns, and as a complement to onsite inspections, we propose the use of virtual proofs of reality for nuclear verification. These proofs could provide a basis for conducting challenge inspections *from a distance* and are designed such that the inspecting party does not lose trust in the remotely used measurement equipment, i.e., inspectors remain confident in the authenticity of the transmitted data.

While the case study of neutron non-irradiation is meant to serve as a simple proof-of-concept for the transferability of virtual proofs to the nuclear domain, there are many possible applications for nuclear safeguards and verification, where remote inspections could be useful. These includes chain-of-custody and other continuity-of-knowledge applications, perimeter control, data commitment, and authentication of measurement equipment. The concept could be particularly valuable for nuclear warhead verification and stockpile monitoring, including for both items and bulk materials, e.g. military fissile material stockpiles and naval fuel.

In the next phase of the project, we plan to demonstrate the virtual proof of neutron non-irradiation experimentally and to examine the viability of other physical unclonable functions, both electronic and non-electronic, for use in radiation measurements. In addition to standard gamma and neutron measurements, there could also be non-nuclear

techniques relevant for nuclear verification such as eddy-current measurements (for chain-of-custody applications)[18] and unique identifiers using reflective particle tags.[19] If demonstrated successfully, virtual proofs of reality could address some of the most difficult remaining challenges for nuclear arms control and disarmament verification.

Acknowledgement. *The authors thank R. J. Goldston and F. d'Errico for their comments on the virtual proof of non-irradiation. Ulrich Rührmair was supported by ERC Project ERCC (FP7/615074).*

Final version, July 15,2016

Endnotes

[1]The unsuccessful negotiations in the 1960s on a Threshold Test Ban Treaty are a relevant example. The United States was seeking 12 to 20 annual inspections in the Soviet Union, when the Soviet Union would not agree on more than 3. Both parties also disagreed on the number of "black box" systems that would need to be installed in each country to monitor seismic activities. See: NAS. (National Academy of Sciences) 1985. Nuclear Arms Control: Background and Issues. Washington, DC: National Academy Press.

[2]S. Philippe, B. Barak, and A. Glaser, Designing protocols for nuclear warhead verification. In Proceedings of the 56th Annual INMM Meeting, Indian Wells, CA (2015).

[3]Remote monitoring is used, for example, to support NPT safeguards and was also discussed as part of SALT II negotiations to remotely monitor ICBMs, see G. J. Simmons, "The history of subliminal channels." IEEE Journal on Selected Areas in Communications 16, no. 4 (1998): 452-462.

[4]G. J. Simmons, "How to insure that data acquired to verify treaty compliance are trustworthy," Proceedings of the IEEE 76, 5 (1988), pp. 621–627.

[5]A recent system based on keys include the *Red Box,* see J. Benz, J. Tanner, and K. Tolk, *Novel Authentication of Monitoring Data Through the Use of Secret and Public Cryptographic Keys,* PNNL-SA-103814, 2014.

[6]U. Rührmair, J. L. Martinez-Hurtado, X. Xu, C. Kraeh, C. Hilgers, D. Kononchuk, J. J. Finley, and W. P. Burleson. "Virtual proofs of reality and their physical implementation." In 2015 IEEE Symposium on Security and Privacy, pp. 70–85. IEEE, 2015.

[7]This section draws heavily on Rührmair, U. et al. (2015), *op. cit.*

[8]O. Goldreich, *Foundations of Cryptography*, 1st ed. Vol. 1, Cambridge: Cambridge University Press, 2001.

[9]R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, "Physical one-way functions," Science 297, no. 5589 (2002): 2026-2030.

[10]Q. Chen, G. Csaba, P. Lugli, U. Schlichtmann, and U. Rührmair. "The bistable ring puf: A new architecture for strong physical unclonable functions." In Hardware-Oriented Security and Trust (HOST), 2011 IEEE International Symposium on, pp. 134-141. IEEE, 2011.

[11]The data and source code are available through the authors upon demand.

[12]R. Pappu et al., *Science, op. cit.*

[13]U. Rührmair, Christian Hilgers, S. Urban, A. Weiershäuser, E. Dinter, B. Forster, and C. Jirauschek. Optical PUFs Reloaded. IACR Cryptology ePrint Archive, Report 2013/215, 2013.

[14]A. Gonzales et al. "Reflective Particle Tag for Arms Control and Safeguards Authentication"." In Proceedings of the 50th INMM Annual Meeting, Tucson AZ, USA. 2009.

[15]S. Philippe, R. J. Goldston, A. Glaser, and F. d'Errico. "A physical zero-knowledge object comparison system for nuclear warhead verification." arXiv preprint arXiv:1602.07717 (2016).

[16]F. d'Errico, "Radiation Dosimetry and Spectrometry with Superheated Emulsions," Nuclear Instruments and Methods in Physics Research B, 184 (2001), 229-254.

[17]S. Philippe, "Safeguarding the Military Naval Nuclear Fuel Cycle," Journal of Nuclear Materials Management, 42.3 (2014)

[18]K. J. Bunch, M. Jones, P. Ramuhalli, J. Benz, L. Schmidt Denlinger, "Supporting Technology for Chain of Custody of Nuclear Weapons and Materials Throughout the Dismantlement and Disposition Processes," *Science & Global Security,* 22 (2), 2014, pp. 111–134.

[19]H. A. Smartt et al., "Status of Non-contact Handheld Imager for Reflective Particle Tags," *55th Annual INMM Meeting,* Atlanta, GA, July 2014.