
Source: U.S. Department of Defense


Source: Google


Source: Paul Shambroom


Source: Microsoft

HOW I LEARNED TO STOP WORRYING AND VERIFY THE BOMB

WHAT THE NUCLEAR WORLD CAN TEACH US ABOUT AI VERIFICATION

Alexander Glaser

Program on Science and Global Security
Princeton University

WMD-AI Verification Exchange

The Institute, San Francisco, California, October 30, 2026
Revision 0.2

Why Nuclear Verification Is So Hard

(and what to do about it)



Mk21 reentry vehicles and containerized W87 warheads at F. E. Warren Air Force Base, Cheyenne, Wyoming, October 1992
Source: Paul Shambroom, paulshambroom.com

DEALING WITH SECRECY



PROTECTING THE SECRET

Many government-sponsored efforts have focused on development of dedicated hardware

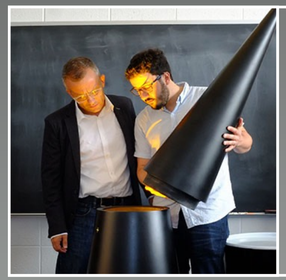
For example, information barriers for warhead confirmation measurements



REINVENTING THE PROBLEM

Never acquire sensitive information to begin with; explore (radically) different technologies and approaches

For example, confirming the absence of a treaty accountable item at a given location



REVEALING THE SECRET

Requirement to protect sensitive information is typically the main reason for complexity of the task

For example, mass and isotopics of fissile material in a nuclear weapon

Source: Author (top and bottom) and Ruhr-Universität Bochum (middle)

Reinventing the Problem

(something old)

SHIFTING THE PARADIGM

2014

ARTICLE

doi:10.1038/nature13457

A zero-knowledge protocol for nuclear warhead verification

Alexander Glaser[1], Boaz Barak[2] & Robert J. Goldston[3]

The verification of nuclear warheads for arms control involves a paradox: international inspectors will have to gain high confidence in the authenticity of submitted items while learning nothing about them. Proposed inspection systems featuring 'information barriers', designed to hide measurements stored in electronic systems, are at risk of tampering and snooping. Here we show the viability of a fundamentally new approach to nuclear warhead verification that incorporates a zero-knowledge protocol, which is designed in such a way that sensitive information is never measured and so does not need to be hidden. We interrogate submitted items with energetic neutrons, making, in effect, differential measurements of both neutron transmission and emission. Calculations for scenarios in which material is diverted from a test object show that a high degree of discrimination can be achieved while revealing zero information. Our ideas for a physical zero-knowledge system could have applications beyond the context of nuclear disarmament. The proposed technique suggests a way to perform comparisons or computations on personal or confidential data without measuring the data in the first place.

Existing nuclear arms-control agreements between the United States and Russia place limits on the number of deployed strategic nuclear weapons. Verification of these agreements takes advantage of the fact that deployed weapons are associated with unique and easily accountable delivery platforms, that is, missile silos, submarines and strategic bombers, to which agreed numbers of warheads are attributed. The next round of nuclear arms-control agreements, however, may place limits on the total number of nuclear weapons and warheads in the arsenals. This would include tactical weapons as well as deployed and non-deployed weapons. Such agreements would require new verification approaches, including inspections of individual nuclear warheads in storage and warheads entering the dismantlement queue. This is a qualitatively new challenge because the design of nuclear weapons is highly classified information that cannot be exposed to international inspectors. A viable verification approach therefore has to resolve the tension between reliably verifying that the inspected warhead is authentic while avoiding disclosure of information about its design[1–4].

Practitioners and policy makers have been well aware of this conundrum, and prior work by national laboratories in the United States, Russia and the United Kingdom addressed it by using 'information barriers'[2,4]. These barriers consist of sophisticated automated systems that process highly classified information measured during an inspection, but only display results in a yes/no manner. Such systems are inherently complex, and require both parties to trust that they have no 'trapdoors' hidden from the inspector, which could be used to cause a system to declare invalid objects as authentic, nor side channels unknown to the host, which could leak classified information to the inspector or others. These concerns are serious obstacles to adopting such systems.

In this work we consider a fundamentally different approach to this problem. Rather than trying to acquire and analyse classified data behind an engineered information barrier, we use the cryptographic notion of zero-knowledge proofs to ensure that sensitive data are never measured in the first place.

Zero-knowledge proofs (with marbles)

These proofs, invented in the 1980s by Goldwasser, Micali and Rackoff[5], have become an important tool of modern cryptography. They achieve the paradoxical goal of allowing one to prove that a statement is true without revealing why it is true. Such proofs are extremely useful for many digital applications, including privacy-preserving data mining, electronic voting and online auctions[6]. To achieve zero knowledge, Goldwasser *et al.* extended the traditional notion of a proof from a static text to a protocol, which involves randomization and interaction between the prover and verifier. At the end of the protocol, the verifier has a high degree of confidence that the statement is correct, while the prover is guaranteed that the verifier did not learn anything about the data underlying the truth of the statement. For our application, the host submitting warheads for inspection takes the role of the prover and the inspector the role of the verifier.

Whereas classical zero-knowledge proofs are digital protocols, proving statements about mathematical objects, we illustrate the concept using a physical zero-knowledge protocol that is closely related to our proposed verification approach (Fig. 1):

Alice (the host) has two small cups both containing X marbles, where X is some number between 1 and 100. She wants to prove to Bob (the inspector) that both cups contain the same number of marbles, without revealing to him what this number X is. To do so, Alice prepares two buckets, which she claims each contain $(100 - X)$ marbles. Bob now randomly chooses into which bucket which cup is poured. Once this is done, Bob verifies that both buckets contain 100 marbles.

The protocol reveals no information on X because, regardless of the value of X, Bob always sees 100 marbles in both buckets. However, if the cups did not have the same number of marbles, then no matter how Alice prepares the buckets, with a probability of 50% after the pouring at least one of the buckets will not contain 100 marbles. If Alice and Bob repeat this game, say, five times, then if Alice consistently cheats she will be caught with probability $(1 - 2^{-5}) > 95\%$.

[1]Department of Mechanical and Aerospace Engineering, Princeton University, E-Quad, Olden Street, Princeton, New Jersey 08544, USA. [2]Microsoft Research, 1 Memorial Drive, Cambridge, Massachusetts 02142, USA. [3]Princeton Plasma Physics Laboratory, Princeton University, PO Box 451, MS 41, Princeton, New Jersey 08540, USA.

©2014 Macmillan Publishers Limited. All rights reserved

2015

SANDIA REPORT
SAND2015-5075
Unlimited Release
June 2015

Report on a Zero-Knowledge Protocol Tabletop Exercise

Peter Marleau,
Erik Brubaker,
Sharon Deland,
Nathan Hilton,
Michael McDaniel,
Richard Schroeppel,
Kevin Seager,
Mary Clare Stoddard

Sandia National Laboratories

Duncan MacArthur
Los Alamos National Laboratory

Prepared by Kevin Seager and Peter Marleau
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

Sandia National Laboratories

2016

THE NEW YORKER

ANNALS OF TECHNOLOGY

THE VIRTUES OF NUCLEAR IGNORANCE

By Alex Wellerstein

September 20, 2016



Alexander Glaser (left) and Sébastien Philippe, of Princeton University, pose with a mockup of a reëntry vehicle, the portion of a nuclear missile that contains the warhead. Photograph courtesy Alexander Glaser and Sébastien Philippe / Nuclear Futures Laboratory

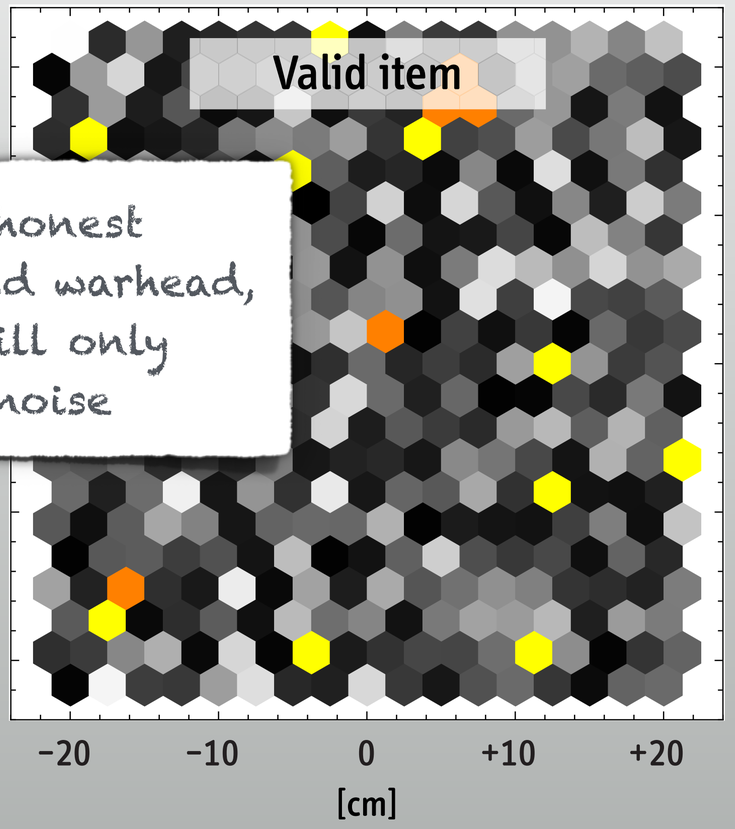
Americans of a certain age are well acquainted with one of Ronald Reagan's pithier maxims: "Trust, but verify." It is a translation of a Russian proverb, and Reagan used it often in his dealings with the Soviet Union over nuclear disarmament, as both countries chipped away tentatively at their Cold War stockpiles. For him, the saying encapsulated the idea that a good treaty would require rigorous inspections, to insure that neither side was dragging its feet or, worse, disposing of decoy bombs, rather than real ones. But how should the process of verification work? If a U.S. scientist were allowed to dismantle a Russian warhead personally, he or she could identify it as legitimate fairly easily, even if it differed in design from American models. But, in doing so, the scientist would necessarily learn many classified details about the weapon, something that the Russians would never consent to. (Nor would the Americans, if their roles were swapped.)

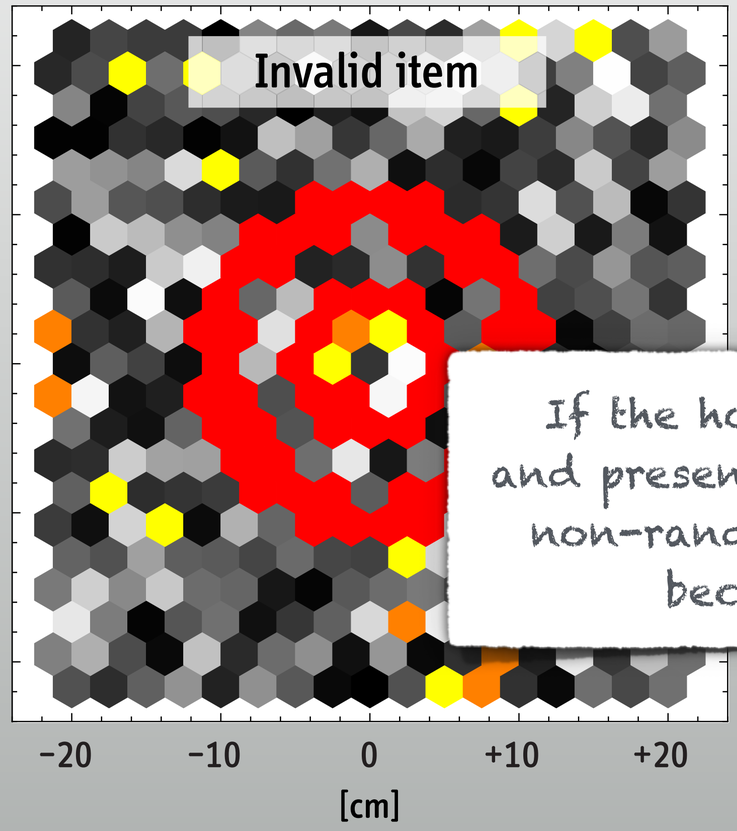
Alexander Glaser, Boaz Barak, and Robert J. Goldston, "A Zero-knowledge Protocol for Nuclear Warhead Verification," Nature, 510, 26 June 2014

IMAGING A NUCLEAR WARHEAD

WITHOUT LEARNING ANYTHING ABOUT ITS DESIGN

If the host is honest and presents a valid warhead, the inspector will only see random noise





If the host tries to cheat and presents a fake warhead, non-random patterns will become visible

Alexander Glaser, Boaz Barak, and Robert J. Goldston, "A Zero-knowledge Protocol for Nuclear Warhead Verification," Nature, 510, 26 June 2014

Guiding Principles

for Good Verification Technologies

SOME (PROPOSED) GUIDING PRINCIPLES

FOR GOOD VERIFICATION TECHNOLOGIES



PURSUE SIMPLICITY AND NON-INTRUSIVENESS

For example, demonstrating the absence of something is more straightforward than confirming the exact nature of a classified item or sensitive activity



OFFER ON-RAMPS

Future agreements may involve partners that are less familiar with bilateral or multilateral inspections; importance of on-ramps, i.e., approaches and procedures that do not require major investments or commitments upfront — but can provide a foundation to build on later

Source: Pavel Podvig (top) and Sascha Kreklau (bottom)

Reinventing the Problem

(something new)

REINVENTING ONSITE INSPECTIONS



PROS & CONS OF ONSITE INSPECTIONS

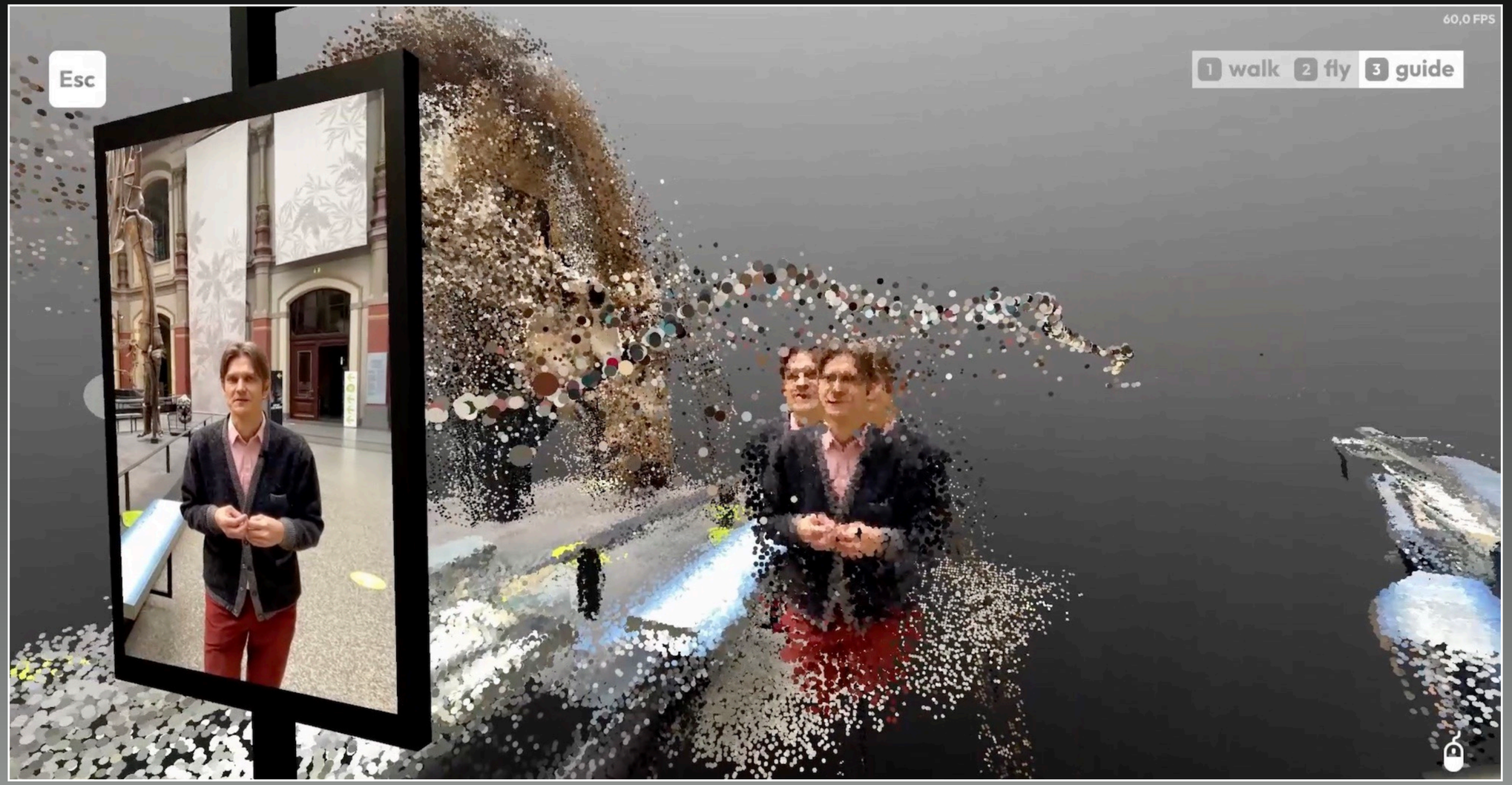
Onsite inspections remain the “gold standard” for IAEA safeguards and nuclear arms-control verification; inspections tend to be costly and are often considered intrusive, especially in an arms-control context



CAN WE (PHYSICALLY) “SEPARATE” HOST & INSPECTOR?

Many concerns could be addressed and resolved if inspectors were not “physically” present onsite; development of the concept could also facilitate engagement of new partners and serve as a catalyst for discussions on the scope of future agreements and their verification

Source: NuDiVe (top) and microsoft.com (bottom)



Jussi Ängeslevä, ART+COM, artcom.de/en/?project=nuform-2020-2022


UREU
211599
UREU
212557


Google
Sweden
049

