Article:
Smith, P.R., Marangon, D. G., Lucamarini, Marco orcid.org/0000-0002-7351-4622 et al. (2 more authors) (2021) Out-of-Band Electromagnetic Injection Attack on a Quantum Random Number Generator. Physical Review Applied. 044044. ISSN 2331-7019

https://doi.org/10.1103/PhysRevApplied.15.044044

Out-of-Band Electromagnetic Injection Attack on a Quantum Random-Number Generator

P. R. Smith,[1,2] D. G. Marangon,[1,*] M. Lucamarini,[1,3] Z. L. Yuan,[1] and A. J. Shields[1]

[1] *Toshiba Europe Ltd, 208 Cambridge Science Park, Milton Road, Cambridge, CB4 0GZ, United Kingdom*

[2] *Cambridge University Engineering Department, 9 JJ Thomson Avenue, Cambridge, CB3 0FA, United Kingdom*

[3] *Department of Physics and York Centre for Quantum Technologies, University of York, YO10 5DD York, United Kingdom*

Random number generators underpin the security of current and future cryptographic systems and are therefore a likely target for attackers. Quantum random number generators have been hailed as the ultimate sources of randomness. However, as shown in this work, the susceptibility of the sensitive electronics required to implement such devices poses a serious threat to their security. We present an out-of-band electromagnetic injection attack on a photonic quantum random number generator through which an adversary can gain full control of the output. In our first experiment, the adversary forces the binary output of the generator to become an alternating string of 1s and 0s, with near 100% success. This attack may be spotted by a vigilant user performing statistical tests on their output strings. We therefore envisage a second more subtle attack in which the adversary forces the output to be a random pattern known to them, thus rendering any protection based on statistical tests ineffective.

I. INTRODUCTION

Random number generators (RNGs) are essential for a wide variety of applications, from lotteries to statistics, from computer simulations to cryptography [1, 2]. For some applications, e.g. computer simulations, the RNG output is only required to be statistically random whereas for others, like cryptography, it is critical that the RNG output is also unpredictable. This guarantees that an adversary cannot steal personal, financial or classified data by predicting or covertly controlling the output of the encryption system. Unpredictable RNGs also underpin the security of quantum key distribution, which provides quantum-based protection to optical telecommunications [3–5]. The generation rate of quantum key distribution can decrease dramatically if the RNG output features even a small imperfection, becoming partially known to the adversary [6, 7].

The necessity for unpredictable random numbers has led to a colossal amount of research into 'physical RNGs', whose randomness is based on physical processes from thermal noise to radioactive decay. Among these, 'photonic RNGs' hold a special place due to the rich variety of implementations they enable, from chaotic lasers to single photon sources, and the promise of integration on chip. In response to our growing reliance on physical RNGs, international standards such as FIPS, NIST SP 800-90B and AIS.31 [8, 9] have been established to guarantee the security of cryptography-orientated RNGs.

Security-wise, it has been shown that the randomness of ring oscillator based RNGs can degrade if their circuits unintentionally act as receiving antennas and pick up electromagnetic radiation from the surrounding environment [10–15]. This undesired behavior can be turned into an attack. In this case, the attack targets the source of randomness itself by locking the ring oscillators to the injected signal. Such attacks, where an adversary injects signals other than those intended to be detected to alter the value of the output, are generally referred to as 'out-of-band signal injection attacks' [16]. These attacks are particularly dangerous because they can be executed remotely and often target the connection between the sensor and the analog-to-digital converter (ADC), which fundamentally cannot be authenticated [16]. They are distinct from high power attacks aimed at disrupting, jamming or burning the victim's system [17], or fault injection attacks targeting digital electronics in cryptographic systems [18, 19], or even side-channel attacks based on physical leakage from the devices [20, 21]. Out-of-band signal injection attacks have been demonstrated on RNGs based on ring oscillators [22–28], medical implants [29] and drones [30], among others [16]. However, there is no study yet of their effectiveness against a quantum device.

As a special subset of physical RNGs, quantum RNGs (QRNGs) provide randomness from a physical process that is fundamentally quantum. The unpredictability of their output is guaranteed by the laws of quantum mechanics, provided that their implementation meets the assumptions made in their theoretical analysis. These assumptions typically identify a security perimeter that the adversary, Eve, cannot cross. Eve can still have full knowledge of the non-quantum characteristics of the devices within the perimeter, but cannot actively exploit them to make her attack more effective. As such, it is assumed that the QRNG is operated in a static environment, perfectly shielded from external signals [1, 2]. Such assumptions are hard to justify in practice. As we find out, in the absence of sufficient shielding, an adversary can control the QRNG output through the unintentional antenna behavior of its components.

* Corresponding author: davide.marangon@crl.toshiba.co.uk

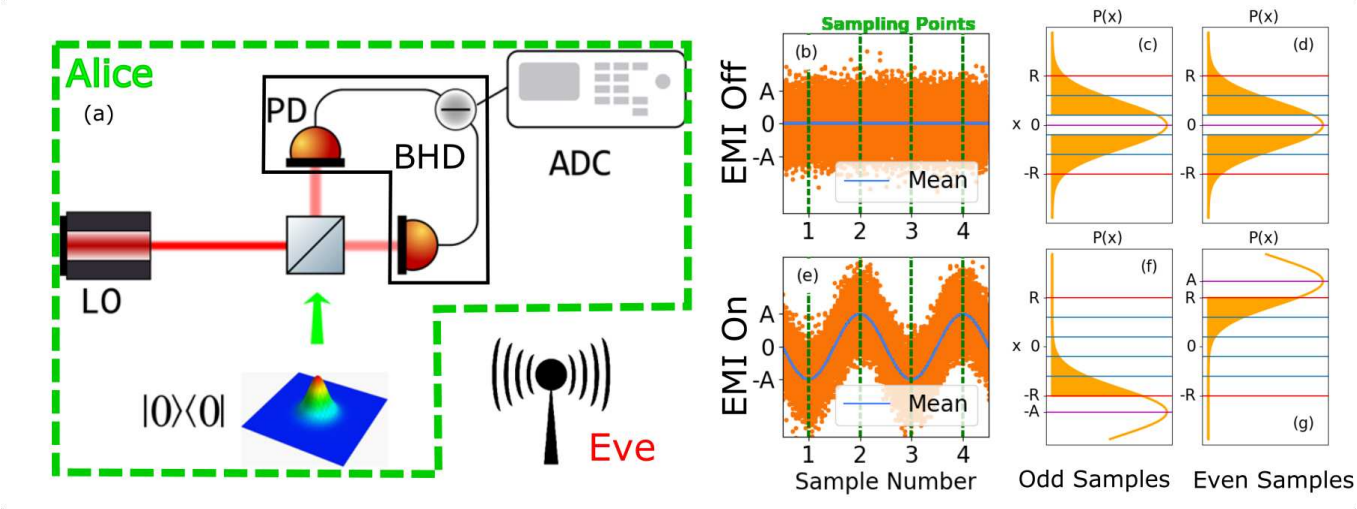

FIG. 1. Models for QRNG and EMI attack. (a) Schematic of a typical CV-QRNG setup used by the user Alice to generate random numbers from measurements of the quadratures of the vacuum field. The setup consists of a local oscillator (LO), a 50:50 beam splitter, on which the LO interferes with a vacuum state (blind port of the beam splitter), a balanced homodyne detector (BHD), in which the outputs of two photodetectors (PD) are subtracted to remove the common mode from the LO, and an analog-to-digital converter (ADC). The attacker Eve uses a radio frequency transmitter to perform the attack. (b-d) In absence of EMI, the idealized experimental points, sampled at a much higher rate and resolution than the user Alice uses, (orange dots in sub-figure (b)) follow a discretized Gaussian with zero mean, represented as a continuous distribution in sub-figures (c) and (d). Alice samples the waveform in correspondence with the dashed green lines in sub-figure (b) and assigns a number depending on which of her ADC bins (blue lines in sub-figures (c), (d) and later on (f), (g)) the output falls into. Because only few samples fall in the outer bins extending from $\pm R$ to $\pm\infty$ (red lines), Eve's best guess of Alice's numbers is the central bin (unshaded region). The probability that Eve's guess is wrong can be found by integrating over the shaded region. (e-g), When Eve injects a sine wave at half Alice's sampling frequency, it shifts Alice's output, moving the mean values (magenta lines) to $-A$ for odd samples (sub-figure (f)) and $+A$ for even samples (sub-figure (g)). Because Alice is unaware of this change, she will not modify the position of her bins (the blue lines in (f), (g) are the same as those in (c), (d)). The attack, however, makes the output likely to fall in one of the outer bins, thereby greatly decreasing the probability that Eve, who now guesses Alice's output will correspond to this outer bin, is wrong. As can be seen from the shaded region in sub-figures (f) and (g) being much smaller than that in sub-figures (c) and (d).

In the following, we describe an out-of-band attack against a QRNG, namely a continuous-variable (CV-QRNG), which is based on the quantum properties of the vacuum field and its subsequent detection via balanced homodyne detection (BHD), see Fig. 1(a). CV-QRNGs have become popular due to their simple implementation and high generation rate [31–41]. In most recent works, they are implemented using commercially available BHDs, for example the Thorlabs PDB480C-AC [33, 34, 36–41].

Earlier works used custom-made BHD circuits, which were observed to suffer from picking up electromagnetic noise from the environment due to the difficulty in shielding the highly sensitive electronics involved [31, 32, 35]. This noise has a classical origin and is therefore typically assumed to be passively monitored by, and hence known to, Eve. The solution to maintain a high secure generation rate has often been to calibrate the output power spectrum of the generator and generate numbers using only the flat regions of the spectrum, which are free from these large classical noise contributions. However, the electromagnetic background is unlikely to remain the same during the operation of the CV-QRNG, especially if a malicious party is actively trying to control the generator output.

In this paper we show how an attacker can create and then actively exploit an electromagnetic side-channel to control the output of a QRNG whilst remaining undetected. To prove our point, we experimentally demonstrate the attack by targeting a typical CV-QRNG that makes use of the most recent BHD equipment. Our attack is not limited to this setup and could be used against any system susceptible to picking up electromagnetic signals, for example generators based on chaotic semiconductor lasers, which make use of similar components [42, 43].

Our attack is based on electromagnetic injection (EMI) and is represented by the model in Fig. 1(b-g). As shown in Fig. 1(b-d), in the absence of EMI we expect the output of the CV-QRNG to be Gaussian distributed, with zero mean and variance σ^2 given by the sum of the quantum noise σ_Q^2, proportional to the power of the local oscillator (LO), and the electronic noise of the measurement system σ_E^2 [32, 35, 44–47]. The injected electromagnetic signal is superimposed on this output, inducing a shift in the mean of the Gaussian distribution. We assume

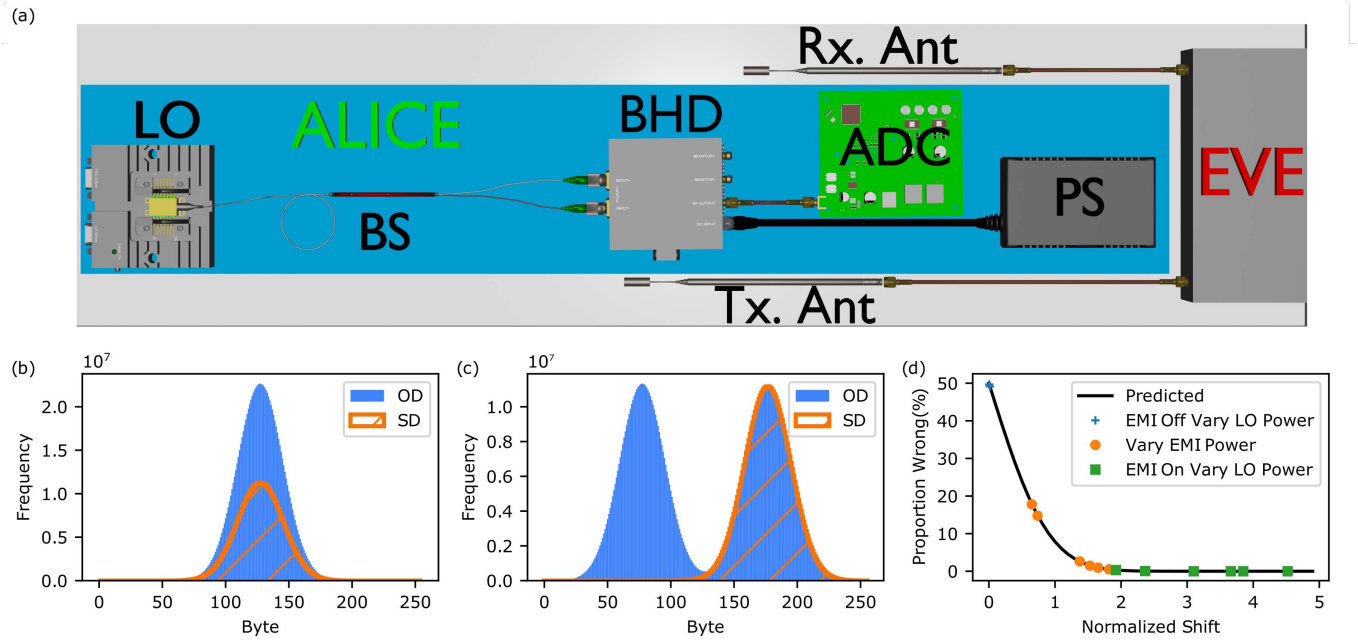

FIG. 2. First experiment on EMI attacks. (a) Schematic of the setup used in the experiment. LO: local oscillator; BS: beam splitter; BHD: balanced homodyne detector; Tx. Ant: transmitting antenna; Rx. Ant: receiving antenna; ADC: analog-to-digital converter; PS: power supply. (b-c) 8 bit histograms of output without (b) and with (c) EMI, with the overall distribution shown in blue and the subsampled distribution, taking every other point, shown in orange. (d) Dependence of the proportion of bits Eve guesses incorrectly on the normalized amplitude of the shift, $\frac{A}{\sqrt{2}\sigma_T}$, she induces in Alice's output.

that variations in the induced shift broaden the shifted Gaussians, increasing their standard deviation to σ_T. In contrast to the aforementioned attacks on ring oscillator based RNGs, this attack targets the hardware between the photodiodes and the ADC rather than the source of randomness itself, which is the vacuum and therefore cannot be degraded through EMI. As illustrated in Fig. 1(e-g), Eve's simplest attack strategy is to inject a sine wave at half Alice's sampling rate, such that the mean of the odd samples is shifted to $-A$ and that of the even samples to $+A$. This results in the overall distribution being double peaked.

In the general case where Alice has an ADC with multiple bins and a finite range R, Eve will aim to shift the distribution such that almost all of Alice's samples fall in the outer bins, which we conservatively assume to extend from $\pm R$ to $\pm\infty$. Given that the magnitude of the induced shift is A, the probability of Eve incorrectly guessing Alice's outcome amounts to

$$P = \begin{cases} \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{R-A}{\sqrt{2}\sigma_T}\right)\right], & \text{if } A < R \\ \frac{1}{2}\mathrm{erfc}\left(\frac{A-R}{\sqrt{2}\sigma_T}\right), & \text{if } A \geq R \end{cases} \quad (1)$$

where $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$ and $\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2}\mathrm{d}t$ (see Appendix A). In the following sections, we focus on the case where Alice has only 2 bins, for which $R = 0$, and present experimental results obtained when Eve is able to synchronize her clock with Alice's remotely. In this case, the autocorrelation and conditional Shannon entropy of the output can also be predicted (see Appendices B&C).

II. RESULTS

A. Injecting Sine Wave

In our first implementation of the EMI attack, the attacker exploits the electromagnetic side-channel in a bidirectional fashion by placing a pair of antennas in the proximity of the QRNG, see Fig. 2(a). With one antenna (Rx), Eve passively picks up Alice's clock and synchronizes with Alice's ADC. With the second antenna (Tx) she actively transmits a sine wave at half Alice's sampling frequency (see Appendix D for further details of the experimental setup). Naturally, Eve can directly attempt an active attack, however, as illustrated in Appendix E, the synchronization greatly improves the attack's success probability.

After setting the phase of her sine signal correctly, Eve expects Alice's output to be an alternating string of 0s and 1s, with the probability of her guessing each bit incorrectly being given by Eq. (1) with $R = 0$. Looking at the histogram of the 8 bit ADC samples in Fig. 2(b-c) we see that, as predicted, Alice's overall output distribution changes from a single Gaussian to a double peaked distribution made up of two Gaussians centered at $\pm A$ when Eve injects her electromagnetic signal. Eve can improve

her control of Alice's output by increasing the amplitude of the induced shift, A, either by placing her transmitting antenna closer to the BHD, or by increasing the transmitted power. The attack would also become more effective if Alice were to reduce her LO power, lowering σ_Q and consequently σ_T. In Fig. 2(d), we show how all these conditions can affect the efficacy of the EMI attack (see Appendix F for further details).

The attack described so far can give Eve full control of Alice's output. However, this version of the attack may be spotted by a vigilant Alice who is performing statistical tests on the ADC samples. The AIS.31 standards for physical generators require that such tests be run continuously to monitor the quality of output randomness [8]. However these tests are computationally intensive and therefore are often not run continuously or at all in most physical generator implementations [13]. Eve could therefore evade this countermeasure by restricting herself to attacking only when Alice is not performing tests, e.g. by monitoring the power drawn by Alice's device in order to ascertain when the tests are being run [20, 48].

Even if Alice were to perform continuous randomness tests, Eve could still attack continuously and remain undetected if she can modify the injected signal to determine the value of each bit at will. She could then send a random sequence known to her which will pass Alice's tests, thus controlling Alice's output whilst rendering her statistical tests ineffective. In the following section we present an implementation of such an attack.

B. Injecting Random Patterns

There are potentially many different schemes which Eve could use to transmit her random pattern to Alice, including variations on frequency shift keying and the use of phased arrays of antennas. We chose to implement a scheme inspired by binary phase shift keying in which the carrier wave, at half Alice's sampling frequency, is multiplied by a non-return-to-zero pattern at Alice's sampling frequency using a mixer. This enables Eve to flip each bit at will (see Appendix D).

This scheme was chosen because it can be implemented with common lab equipment and a single transmitting antenna. In order to force Alice's output to replicate her desired pattern, Eve must ensure that the power spectrum of the CV-QRNG output closely matches that of her transmitted signal for at least one whole sideband. This is challenging to achieve in practice as the RF frequency response of Alice's setup is unlikely to be flat across a whole sideband i.e. her setup is unlikely to be equally good at receiving signals across a whole sideband.

In our implementation, we concentrate on sending strings of repetitive 32 bit patterns, chosen using a pseudo-RNG. Counting the number of matches between the sequence injected by Eve and the measured output for 1 281 250 samples in 3000 different patterns, we find an average match of 69.8% with a standard deviation of

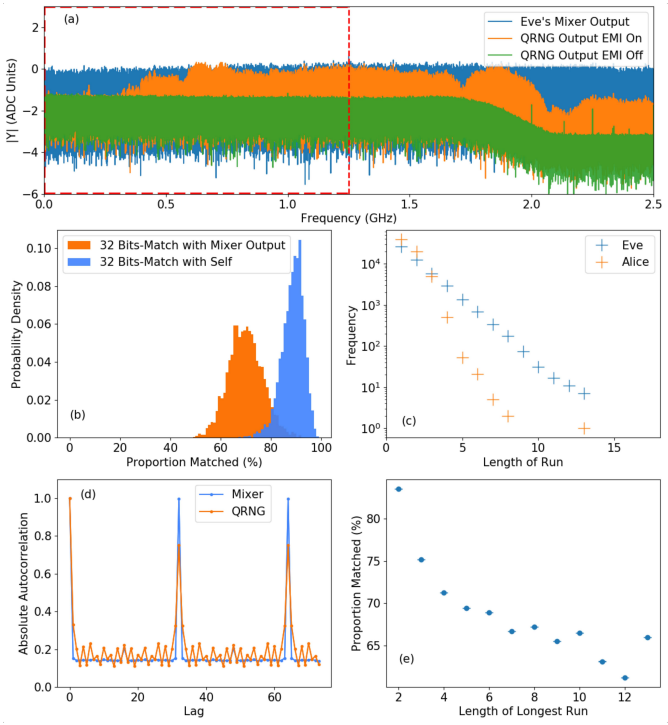

FIG. 3. Injecting random patterns. (a) Comparison between the power spectra of the output from Eve's mixer (blue) and that of the CV-QRNG output in the absence (green) and presence (orange) of EMI when Eve is attempting to send a repeating 8192 bit pattern. The lower sideband of Eve's signal in which she requires the power spectra to match is shown by the red dashed perimeter. (b) Histograms of proportion of bits matched between 3000 32 bit patterns transmitted by Eve and the corresponding outputs from the CV-QRNG (orange), and the match obtained by instead assuming Eve knows the first 32 bits in the CV-QRNG output and comparing this with the remainder of the output (blue). (c) Comparison of occurrence of different run lengths in the patterns sent by Eve and those received by Alice. (d) Mean absolute autocorrelations of mixer and CV-QRNG outputs showing that a repeating 32 bit pattern is present in both. (e) Dependence of proportion matched on length of maximum run present in the transmitted pattern.

6.7%, as shown in Fig. 3(b). This match can be interpreted as the probability of Eve correctly guessing each bit of Alice's output. Eve could extend these patterns to arbitrary lengths in order to better conceal her attack.

Much of the mismatch between Eve's injected sequence and the one obtained by Alice can be attributed to the RF frequency response of Alice's setup. Fig. 3(a), obtained by sending a repeating 8192 bit pattern and sampling the output at 40 GSamples/s using an 8 bit oscilloscope, shows that the CV-QRNG setup is poor at picking up injected signals below 500 MHz, as is clear from the mismatch between the power spectra of the output from Eve's mixer, shown in blue, and that of Alice's CV-QRNG output, in orange. This in turn leads to the loss of the longer runs, i.e. uninterrupted sequences of identical

bits, in Alice's output, as shown in Fig. 3(c), and consequently a lowering of the average match. As highlighted in Fig. 3(e) the match decreases almost linearly with the length of the longest run in the pattern Eve is attempting to transmit. This is due to the loss of the low frequency components in the received signal. Eve ought to be able to correct for this distortion and increase her match if she has prior knowledge of the frequency response, by amplifying the parts of the spectrum for which the frequency response is weaker prior to transmission. Otherwise she could increase her average match by restricting herself to sending patterns containing only short runs.

Despite this imperfect match between Eve's target and the patterns received by Alice, it is clear from the autocorrelation shown in Fig. 3(d) that the received patterns have 32 sample-long repeating patterns within them. If instead of comparing the CV-QRNG output to that of Eve's mixer we compare it to itself, taking the first 32 bits of Alice's output as the pattern, the average match rises to 88.7% with a standard deviation of 4%, as shown in Fig. 3(b). The remaining mismatch is attributed to the lack of power in the injected signal. Therefore Eve's guessing probability would improve considerably if she was aware of how her pattern had been distorted.

Now consider a second scenario in which Alice uses the band from 625 MHz to 1.25 GHz to generate her output. Eve can adapt her input, using the upper sideband rather than the lower sideband to transmit her pattern by setting her carrier wave frequency to 625 MHz and mixing this with a 1.25 GHz pattern. In this case, after filtering our experimental results show that Eve's match increases to 88% on average with a standard deviation of 10% for 3000 32 bit patterns, as shown in Fig. 4(a). This is due to the fact that the RF frequency response of Alice's setup is relatively flat throughout this region, see Fig. 3(a), meaning that Eve is more successful in transmitting her pattern. This is reflected in the fact that the longer runs are preserved in the CV-QRNG output, as shown in Fig. 4(b). If as before we compare Alice's output to the first 32 bits within it the match rises to 94% with a standard deviation of 3%, see Fig. 4(a).

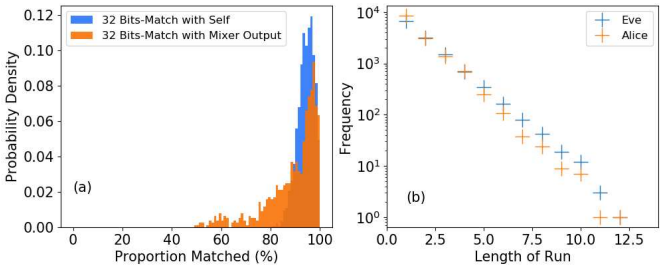

FIG. 4. (a) Proportion matched after filtering when Eve sends 3000 32 bit patterns in the case where Alice is using the band from 625 MHz to 1.25 GHz to generate her output, comparing Alice's output to Eve's injected signal (orange) and to the first 32 bits in Alice's output (blue). (b) Comparison of occurrence of different run lengths in the patterns sent by Eve and received by Alice.

III. CONCLUSION

In this work we presented an out-of-band signal injection attack on a photonic QRNG through which an adversary can gain full control of the output through EMI. We present three proof of principle implementations against a CV-QRNG with a binary output, using common lab equipment and a wideband isotropic antenna. The first was able to achieve near perfect control of the output by exploiting the out-of-band electromagnetic channel in a bidirectional fashion: eavesdropping Alice's clock for synchronization and injecting a sine wave. This attack forces the output to become a series of alternating 1s and 0s and could therefore be spotted by a vigilant user performing statistical tests on their output. We therefore investigate two scenarios in which Eve can achieve high degrees of matching between random patterns chosen by her and the CV-QRNG output. We anticipate these matches could be increased with a more powerful and sophisticated transmitter setup, or perhaps a different modulation scheme. When perfected, such an attack would render any protection based on statistical tests on the output ineffective, highlighting the need for implementing countermeasures specific to EMI attacks (see Appendix H).

Our results are pertinent to the more common issue of unintentional electromagnetic interference, which may lead to the randomness of RNGs being degraded if they are deployed in server racks or other noisy environments with insufficient shielding. Out-of-band signal injection attacks on quantum technologies, such as quantum key distribution systems, are an as yet unexplored research area and could therefore pose previously unidentified security threats that shall be investigated in future works.

ACKNOWLEDGMENTS

P.R.S. gratefully acknowledges financial support from the EPSRC (Award No. 1771797) CDT in Integrated Photonic and Electronics Systems and Toshiba Research Europe, Limited.

APPENDIX A: PROPORTION EVE GUESSES INCORRECTLY

Eve aims to shift Alice's distribution such that the outcome of Alice's measurement is most likely to fall in one of the outer bins of her ADC, which are assumed to extend from $\pm R$ to $\pm\infty$, and guesses that this will be Alice's output. The probability that this will not be the case can be found by integrating the filled region under each curve in Fig. 5. Given that the variance of the Gaussian is σ_T^2, the probability that Eve will guess the outcome of Alice's measurement incorrectly is given by Eq. 1 in the main text.

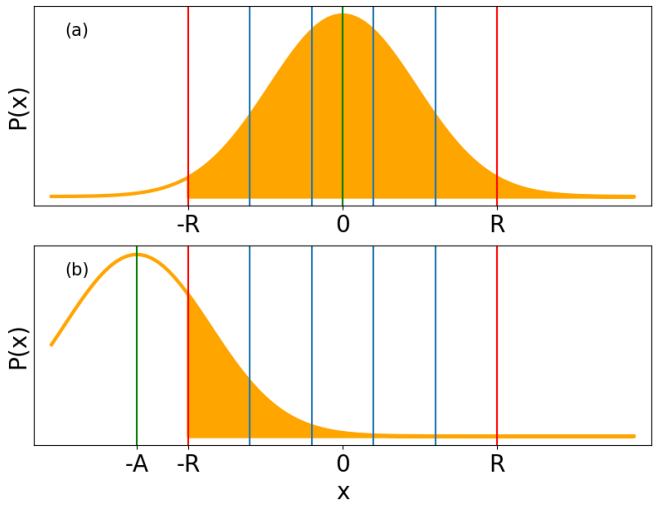

FIG. 5. (a-b) Probability density function of Alice's BHD output in the absence (a) and presence (b) of Eve's electromagnetic signal, with which she shifts the mean of Alice's Gaussian distribution, indicated by the green line, to $-A$. The blue lines indicate the edge of Alice's ADC bins. The red lines indicate the bottom edge of the two outer ADC bins, which are assumed to extend from $\pm R$ to $\pm\infty$. The probability that Eve will guess Alice's output incorrectly can be found by integrating the filled region.

The widths of the inner bins of Alice's ADC are shown as being equal in Fig. 5, but could also be chosen such that the outcome of Alice's measurement is equally likely to fall in each bin, see for example [31]. In either case, Eve can maximize her guessing probability by maximizing the shift A that she induces in Alice's output. Determining how Eve can modulate her injected signal to maximize her guessing probability whilst remaining undetected by statistical tests run by Alice on her output when she has more than 2 bins is an interesting and complex problem, which would depend not only on Alice's choice of binning but also on which statistical tests she is performing, and shall not be discussed in any further detail in this work.

APPENDIX B: ABSOLUTE AUTOCORRELATION

Fig. 6(a) shows that the data goes from being weakly correlated for low lags, due to the finite bandwidth of the detector, in the absence of EMI, to being strongly correlated for all lags when Eve injects a sine wave. The absolute value of the autocorrelation of the binary output for non-zero lags when Eve is injecting her signal can be shown to be given by:

$$r = \left[\mathrm{erf}\left(\frac{A}{\sqrt{2}\sigma_T}\right)\right]^2. \tag{B1}$$

Fig. 6(b) shows that the data obtained fit this prediction well.

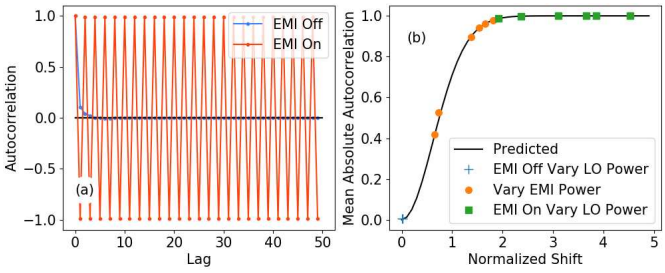

FIG. 6. (a) Autocorrelation of the binary QRNG output in the presence and absence of EMI. (b) Dependence of the mean absolute autocorrelation for non-zero lags on the normalized amplitude of the shift induced by Eve.

The autocovariance function at lag k, for $k \geq 0$ is defined as:

$$s_k = \frac{1}{n}\sum_{i=1}^{n-k}(y_i - \bar{y})(y_{i+k} - \bar{y}) \tag{B2}$$

where the mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{i=n} y_i$. The autocorrelation function at lag k, for $k \geq 0$, is defined as:

$$r_k = \frac{s_k}{s_0} \tag{B3}$$

For the sake of compactness, assume that after thresholding at 0 V Alice's output will be $+1$ if the BHD output was positive and -1 if it was negative, such that $\bar{y} = 0$. We then have:

$$r_k = \frac{\sum_{i=1}^{n-k} y_i y_{i+k}}{\sum_{i=1}^{n} y_i^2} \tag{B4}$$

Taking the limit as $n \to \infty$ and assuming that each element is independent we can rewrite each sum as a sum over possible outcomes weighted by their probabilities. Defining $p_{i,j}$ as the probability that Alice's output will be i given that Eve predicts it to be j, and using Eq. 1 with $R = 0$ we have:

$$p_{1,-1} = p_{-1,1} = \frac{\mathrm{erfc}\left(\frac{A}{\sqrt{2}\sigma_T}\right)}{2} \tag{B5}$$

and $p_{1,1} = p_{-1,-1} = 1 - p_{1,-1}$. Considering the odd and even terms in the sums separately the autocorrelation for odd lags is then given by:

$$r_{\mathrm{odd}} = \frac{p_{1,-1}p_{1,1} - p_{-1,-1}p_{1,1} + p_{-1,-1}p_{-1,1} - p_{1,-1}p_{-1,1}}{2(p_{1,-1} + p_{-1,-1})} + \frac{p_{1,1}p_{1,-1} - p_{-1,1}p_{1,-1} + p_{-1,1}p_{-1,-1} - p_{1,1}p_{-1,-1}}{2(p_{1,1} + p_{-1,1})} \tag{B6}$$

which using the results above simplifies to:

$$\begin{aligned} r_{\mathrm{odd}} &= -(p_{1,1}^2 - 2p_{1,1}p_{1,-1} + p_{1,-1}^2) \\ &= -(p_{1,1} - p_{1,-1})^2 \\ &= -(1 - 2p_{1,-1})^2 \end{aligned} \tag{B7}$$

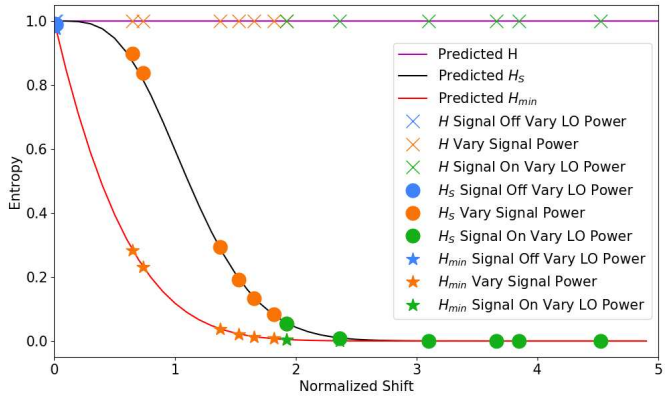

FIG. 7. Comparison between predicted and experimental values of the Shannon entropy (H), conditional Shannon entropy (H_S) and conditional min-entropy (H_{min}) of 1 Gbit binary output from Alice's QRNG as a function of normalized shift imparted by Eve.

Similarly it can be shown that $r_{\text{even}} = +(1-2p_{1,-1})^2$. Substituting in the probability from above, we find that absolute value of the autocorrelation for non-zero lags is given by:

$$r = \left[1 - \text{erfc}\left(\frac{A}{\sqrt{2}\sigma_T}\right)\right]^2 = \left[\text{erf}\left(\frac{A}{\sqrt{2}\sigma_T}\right)\right]^2 \tag{B8}$$

which matches Eq. B1.

APPENDIX C: CONDITIONAL SHANNON ENTROPY

The success of Eve's attack in the case where she injects a sine wave into Alice's QRNG whose output is binary can further be demonstrated by evaluating the conditional Shannon entropy of Alice's output, defined as:

$$H_S = -\sum_y p(y) \sum_x p(x|y) \log_2 p(x|y) \tag{C1}$$

where the conditional probability $p(x|y) = \frac{p(yx)}{p(y)}$ is the probability that event x will occur, given that event y just occurred [49]. In the case of a binary output this can be rewritten as:

$$H_S = -p(00)\log_2\frac{p(00)}{p(0)} - p(01)\log_2\frac{p(01)}{p(0)} - p(10)\log_2\frac{p(10)}{p(1)} - p(11)\log_2\frac{p(11)}{p(1)} \tag{C2}$$

where for example $p(0|1) = p(10)/p(1)$ corresponds to the probability that Alice's next output will a 0 given that her last bit was a 1. Details of the procedure to obtain the necessary probabilities from experimental data can be found in [49].

As shown in Fig. 7 Alice's conditional Shannon entropy decreases as the normalized shift, $\frac{A}{\sqrt{2}\sigma_T}$, that Eve imparts to Alice's BHD output increases, reaching 0 for sufficiently large shifts. If Alice were evaluating the conditional Shannon entropy of her output, it would be clear to her at this point that her output is completely predictable. Crucially if instead Alice was simply evaluating the Shannon entropy of her output $H = -\sum_x p(x)\log_2 p(x)$, the value she would obtain would remain close to 1, and the attack would go unnoticed.

The conditional Shannon entropy of Alice's output can be predicted by calculating each the probabilities in Eq. C2 from Alice's perspective who is assumed to be unaware of Eve's attack:

$$p(0) = p(1) = \frac{1}{2} \tag{C3}$$

$$p(00) = p(11) = \frac{1}{4}\text{erfc}\left(\frac{A}{\sqrt{2}\sigma_T}\right)\left(2 - \text{erfc}\frac{A}{\sqrt{2}\sigma_T}\right) \tag{C4}$$

$$p(01) = p(10) = \frac{1}{8}\left(\text{erfc}\left(\frac{A}{\sqrt{2}\sigma_T}\right)\right)^2 + \frac{1}{8}\left(2 - \text{erfc}\left(\frac{A}{\sqrt{2}\sigma_T}\right)\right)^2 \tag{C5}$$

The conditional Shannon entropy is then given by:

$$H_S = -x\log_2 x - y\log_2 y, \tag{C6}$$

with

$$x = \frac{1}{2}\text{erfc}\left(\frac{A}{\sqrt{2}\sigma_T}\right)\left(2 - \text{erfc}\frac{A}{\sqrt{2}\sigma_T}\right), \tag{C7}$$

$$y = \frac{1}{4}\left[\left(\text{erfc}\frac{A}{\sqrt{2}\sigma_T}\right)^2 + \left(2 - \text{erfc}\frac{A}{\sqrt{2}\sigma_T}\right)^2\right]. \tag{C8}$$

As shown in Fig. 7 the experimental data fits this prediction well. The conditional min-entropy, for which we assume that the side information available to Eve is whether she was trying to send a 1 or a 0 at each sampling point, is also plotted in Fig. 7 to highlight how much this side information improves Eve's guessing probability. This guessing probability would further increase if, as in many QRNG protocols, we assume that the electronic noise from the detector is known to Eve.

APPENDIX D: EXPERIMENTAL METHODS

For our proof of principle experimental implementation we focus on the case in which Alice obtains her digital output by thresholding the BHD output at 0 V. The QRNG setup consists of a laser diode, connected to a

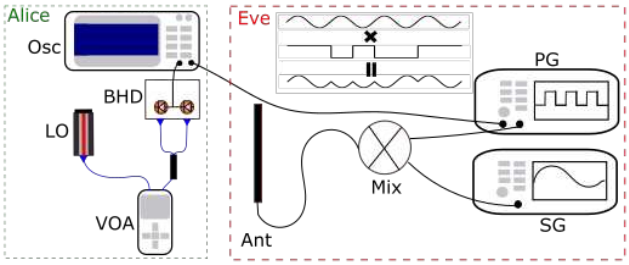

FIG. 8. Schematic of setup. LO: local oscillator; VOA: variable optical attenuator; BHD: balanced homodyne detector; Osc: oscilloscope; PG: pattern generator; SG: signal generator; Mix: mixer; Ant: antenna. Inset: Illustration of Eve's modulation scheme.

variable optical attenuator (VOA), the output of which is connected to a 50:50 fibre coupler. The second input of the coupler is blocked as to provide a vacuum input. The two outputs from the coupler are connected to a Thorlabs PDB480C-AC BHD. Unless otherwise stated, the LO power is set just below the power at which the BHD saturates, such that around 4.7 mW is incident on each photodiode.

For the experiments in which Eve is sending a sine wave, the output is sampled at 1 GSamples/s using a dedicated ADC board. In this case the board's 1 GHz clock is picked up by placing an Aim-TTi PSA-ANT2 antenna close to the ADC board, the output from which is filtered and then frequency divided to provide the 10 MHz reference for Eve's setup, making the attack contactless. A schematic of this setup is shown in Fig. 2(a).

For the experiments in which more complex patterns are sent, the output from the BHD is sampled at 40 GSamples/s using an oscilloscope, then downsampled to 2.5 GSamples/s in post-processing. Sampling at a high rate then downsampling to the required rate gives more flexibility in choosing the sampling point. In this case we assume that Eve has direct access to Alice's clock and trigger the oscilloscope on Eve's pattern generator output. A schematic of the setup is shown in Fig. 8.

The injected electromagnetic signal is generated using a signal generator in the case where we send a sine wave, and with a combination of said signal generator and a pattern generator whose outputs are combined using a mixer when sending more complex patterns. The signal is amplified to 24 dBm and then transmitted using an Aim-TTi PSA-ANT2 isotropic wideband antenna placed a few centimeters away from the BHD.

Whilst we cannot be certain which part of the system acts as an unintentional antenna and picks up the electromagnetic signal we suspect it is a combination of the power supply cable and the output SMA, as placing the antenna parallel and in close proximity to these close to their connections to the BHD box produces the largest response. The BHD circuit board may also be responsible although it is held in a shielded aluminum box [16, 50, 51].

APPENDIX E: UNSYNCHRONIZED CLOCKS

In the main text we considered the case where Eve can obtain Alice's clock and hence synchronize her attack with Alice's sampling, we now consider what happens if this is not the case. To keep the analysis simple we restrict ourselves to considering the case where Eve sends a sine wave at half Alice's sampling frequency. If Eve is unable to synchronize her clock with Alice's, Alice will no longer sample Eve's injected sine wave on the extrema, instead Alice's sampling point will drift along Eve's sine wave. Assuming that the clocks are stable, the shift imparted by Eve will evolve over time, t, as $A\cos(\Delta f t)$, where Δf is the difference between the frequency of Eve's signal and half Alice's sampling rate. As shown by the experimental results in Fig. 9 if we subsample short sections of the Alice's output, taking every other point, and calculate the mean shift imparted by Eve, we see that it oscillates sinusoidally as predicted.

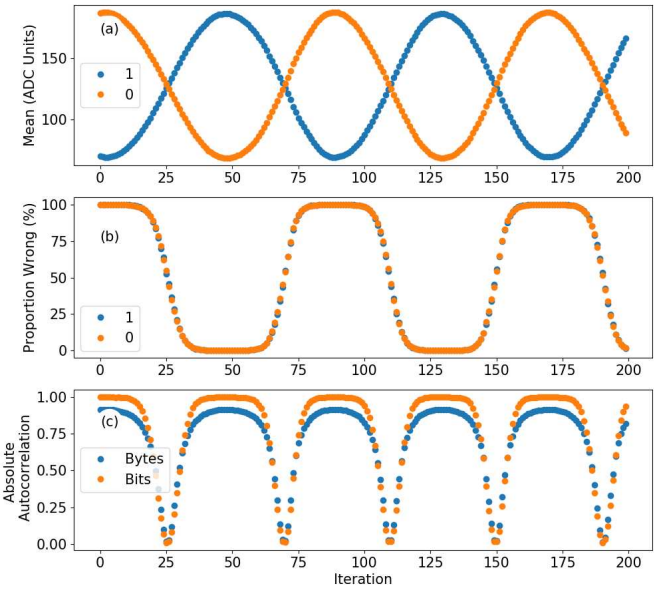

FIG. 9. Evolution over time of (a) the mean shift imparted by Eve, (b) the proportion of bits Eve guesses incorrectly and (c) the absolute value of the non-zero lag autocorrelation for 100,000 samples when Eve and Alice's clocks are not synchronized.

The proportion wrong and absolute non-zero autocorrelation show the same periodicity and can be accurately predicted from the normalized shift imparted by Eve using Eq. 1 and Eq. B1 respectively, as shown in Fig. 10 .

APPENDIX F: EXPERIMENTALLY VARYING THE NORMALIZED SHIFT

Fig. 11 and Fig. 12 provide further details of the parameters used to produce Fig. 2 (d). In Fig. 11(a) we show

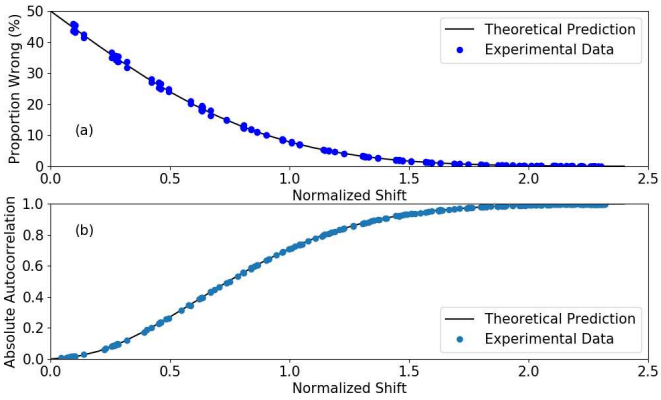

FIG. 10. Proportion of bits that Eve guesses incorrectly (a) and mean absolute non-zero autocorrelation for binary output (b) as a function of the normalized shift imparted by Eve.

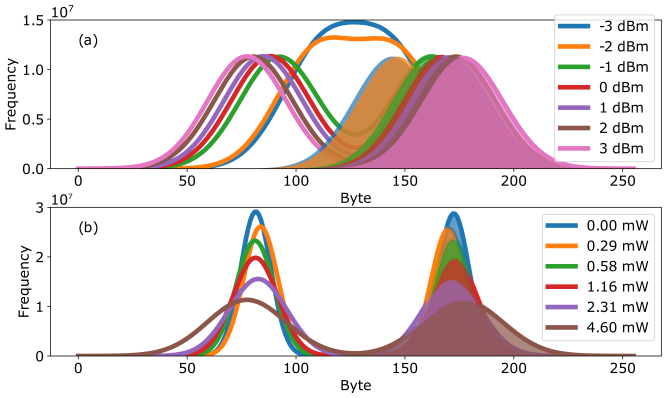

FIG. 11. 8 bit histograms for (a) fixed 4.6 mW LO power and varying signal generator power, and (b) fixed 3 dBm signal generator power and varying LO power. Overall distribution (solid line). Distribution taking every other sample (fill).

that increasing the power output by Eve's signal generator increases the shift imparted by Eve and hence the distance between the two extrema in the overall distribution(solid line). The distributions obtained after subsampling, taking every other point, are also plotted (filled) to emphasise the fact that they are simply shifted Gaussians.

Fig. 11(b) shows that increasing the LO power increases the width of the Gaussians. Further to this Fig. 12(a) shows that the variance of the subsampled distributions remains proportional to the LO power in the presence of EMI and that Eve's attack does not significantly change this variance compared to that obtained when she is not attacking. Fig. 12(b) shows the dependence of the amplitude of the shift imparted by Eve on the power output from Eve's signal generator.

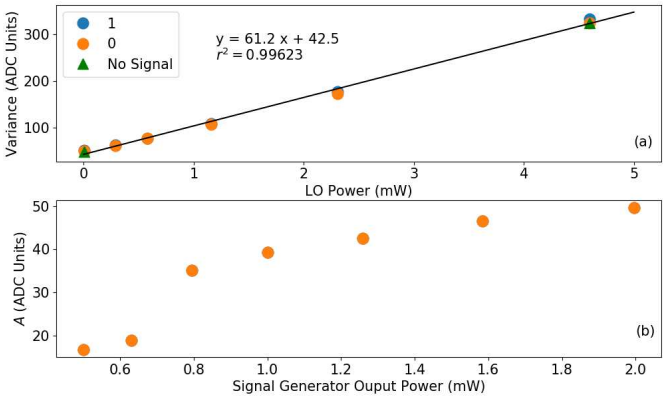

FIG. 12. (a) Dependence of the variance of the subsampled waveforms, taking every other sample, on the LO power . (b) Dependence of the amplitude of the shift, A, imparted by Eve on the signal generator power.

APPENDIX G: RANDOMNESS EXTRACTION

It is worth pointing out that CV-QRNGs are normally provided with a unit for the application of so-called randomness extractors, i.e., algorithms to enhance the statistical uniformity of ADC samples and make them more difficult to predict. However, with this kind of attack this unit would be of little use. All the post processing steps applied by Alice in order to extract her final output from the bits must be known to Eve [52] and do not add any entropy to the output, therefore they cannot make the final output unpredictable to Eve if she knows the raw input. Worse still, it is common to only apply randomness test to the post-processed bits in CV-QRNG implementations [32, 34, 36–38, 40, 41, 44–47]. Such tests ought to be passed even in the case where Alice's raw output is a string of alternating 0s and 1s if for example the raw output is hashed using a Toeplitz matrix, meaning that Alice will fail to spot even this simpler version of our attack [53].

APPENDIX H: COUNTERMEASURES

Countermeasures against electromagnetic interference, intentional or not, have been the subject of extensive research. It has been shown that shielding, differential coupling and filtering can be applied to effectively attenuate electromagnetic signals [16, 29, 54, 55]. Such countermeasures only attenuate Eve's signal and can therefore be overcome by Eve sending a more powerful signal. If instead Alice wishes to detect that the attack is taking place, she could monitor the power reaching the ADC as this will increase considerably during the attack. Alice may also place an antenna close to the RNG to monitor the electromagnetic background [29]. The need to implement monitoring ahead of the ADC has previously been highlighted in [56, 57] and in the AIS.31 standards in the

form of total failure tests on the entropy source [8].

As discussed above, in the case where Alice has only 2 bins, it is possible for Eve to adapt her input to render Alice's statistical tests ineffective. This becomes more difficult for Eve as the number of ADC bins increases as any shift imparted to Alice's output by Eve will increase the occurrence of samples in the outer bins. Alice may then be able to detect the attack by counting the occurrence of samples in these bins. This countermeasure can be made more effective if Alice counts the number of samples in the outer bins after randomly switching off the LO, as in a typical CV-QRNG setup this will drastically reduce the probability of a measurement falling in the outer bins [32, 55, 58]. A detailed overview of further potential countermeasures against out-of-band signal injection attacks in general can be found in [16, 51].

[1] X. Ma, X. Yuan, Z. Cao, B. Qi, and Z. Zhang, Quantum random number generation, npj Quantum Inf. **2**, 16021 (2016).

[2] M. Herrero-Collantes and J. C. Garcia-Escartin, Quantum random number generators, Rev. Mod. Phys. **89**, 015004 (2017).

[3] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Quantum cryptography, Rev. Mod. Phys. **74**, 145 (2002).

[4] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, The security of practical quantum key distribution, Rev. Mod. Phys. **81**, 1301 (2009).

[5] E. Diamanti, H.-K. Lo, B. Qi, and Z. Yuan, Practical challenges in quantum key distribution, npj Quantum Inf. **2**, 16025 (2016).

[6] J. Bouda, M. Pivoluska, M. Plesch, and C. Wilmott, Weak randomness seriously limits the security of quantum key distribution, Phys. Rev. A **86**, 062308 (2012).

[7] H.-W. Li, Z.-M. Xu, and Q.-Y. Cai, Small imperfect randomness restricts security of quantum key distribution, Phys. Rev. A **98**, 062325 (2018).

[8] W. Killmann and W. Schindler, A proposal for: Functionality classes for random number generators, ser. BDI, Bonn (2011).

[9] M. S. Turan, E. Barker, J. Kelsey, K. McKay, M. Baish, and M. Boyle, Nist special publication 800-90b: recommendation for the entropy sources used for random bit generation, january 2016.

[10] P. Bayon, L. Bossuet, A. Aubert, V. Fischer, F. Poucheret, B. Robisson, and P. Maurine, Contactless electromagnetic active attack on ring oscillator based true random number generator, in *International Workshop on Constructive Side-Channel Analysis and Secure Design* (Springer, 2012) pp. 151–166.

[11] S. Buchovecka and J. Hlaváč, Frequency injection attack on a random number generator, in *2013 IEEE 16th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS)* (IEEE, 2013) pp. 128–130.

[12] P. Maistri, R. Leveugle, L. Bossuet, A. Aubert, V. Fischer, B. Robisson, N. Moro, P. Maurine, J.-M. Dutertre, and M. Lisart, Electromagnetic analysis and fault injection onto secure circuits, in *2014 22nd International Conference on Very Large Scale Integration (VLSI-SoC)* (IEEE, 2014) pp. 1–6.

[13] P. Bayon, L. Bossuet, A. Aubert, and V. Fischer, Fault model of electromagnetic attacks targeting ring oscillator-based true random number generators, Journal of Cryptographic Engineering **6**, 61 (2016).

[14] M. Madau, M. Agoyan, J. Balasch, M. Grujić, P. Haddad, P. Maurine, V. Rožić, D. Singelée, B. Yang, and I. Verbauwhede, The impact of pulsed electromagnetic fault injection on true random number generators, in *2018 Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC)* (IEEE, 2018) pp. 43–48.

[15] S. Osuka, D. Fujimoto, Y.-i. Hayashi, N. Homma, A. Beckers, J. Balasch, B. Gierlichs, and I. Verbauwhede, Em information security threats against ro-based trngs: The frequency injection attack based on iemi and em information leakage, IEEE T. Electromagn. C. (2018).

[16] I. Giechaskiel and K. Rasmussen, Taxonomy and challenges of out-of-band signal injection attacks and defenses, IEEE Commun. Surv. Tutor. **22**, 645 (2019).

[17] F. Sabath, What can be learned from documented intentional electromagnetic interference (iemi) attacks?, in *2011 XXXth URSI General Assembly and Scientific Symposium* (IEEE, 2011) pp. 1–4.

[18] A. Barenghi, L. Breveglieri, I. Koren, and D. Naccache, Fault injection attacks on cryptographic devices: Theory, practice, and countermeasures, Proceedings of the IEEE **100**, 3056 (2012).

[19] D. Karaklajić, J.-M. Schmidt, and I. Verbauwhede, Hardware designer's guide to fault attacks, IEEE. T. VLSI Syst. **21**, 2295 (2013).

[20] P. Kocher, J. Jaffe, and B. Jun, Differential power analysis, in *Annual International Cryptology Conference* (Springer, 1999) pp. 388–397.

[21] D. J. Bernstein, Is the security of quantum cryptography guaranteed by the laws of physics?, arXiv preprint arXiv:1803.04520 (2018).

[22] M. Šimka and P. Komenského, Active non-invasive attack on true random number generator, in *6th PhD Student Conference and Scientific and Technical Competition of Students of FEI TU Košice, Košice, Slovakia* (Citeseer, 2006) pp. 129–130.

[23] A. T. Markettos and S. W. Moore, The frequency injection attack on ring-oscillator-based true random number generators, in *International Workshop on Cryptographic Hardware and Embedded Systems* (Springer, 2009) pp. 317–331.

[24] N. Bochard, F. Bernard, V. Fischer, and B. Valtchanov, True-randomness and pseudo-randomness in ring oscillator-based true random number generators, International Journal of Reconfigurable Computing **2010** (2010).

[25] M. Soucarros, C. Canovas-Dumas, J. Clédière, P. Elbaz-Vincent, and D. Réal, Influence of the temperature on true random number generators, in *2011 IEEE International Symposium on Hardware-Oriented Security and Trust* (IEEE, 2011) pp. 24–27.

[26] H. Martin, T. Korak, E. San Millán, and M. Hutter, Fault attacks on strngs: Impact of glitches, temperature,

and underpowering on randomness, IEEE T. Inf. Foren. Sec. **10**, 266 (2014).

[27] Y. Cao, V. Rožić, B. Yang, J. Balasch, and I. Verbauwhede, Exploring active manipulation attacks on the tero random number generator, in *2016 IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS)* (IEEE, 2016) pp. 1–4.

[28] H. Martin, P. Martin-Holgado, P. Peris-Lopez, Y. Morilla, and L. Entrena, On the entropy of oscillator-based true random number generators under ionizing radiation, Entropy **20**, 513 (2018).

[29] D. F. Kune, J. Backes, S. S. Clark, D. Kramer, M. Reynolds, K. Fu, Y. Kim, and W. Xu, Ghost talk: Mitigating emi signal injection attacks against analog sensors, in *2013 IEEE Symposium on Security and Privacy* (IEEE, 2013) pp. 145–159.

[30] Y. Son, H. Shin, D. Kim, Y. Park, J. Noh, K. Choi, J. Choi, and Y. Kim, Rocking drones with intentional sound noise on gyroscopic sensors, in *24th {USENIX} Security Symposium ({USENIX} Security 15)* (2015) pp. 881–896.

[31] T. Symul, S. Assad, and P. K. Lam, Real time demonstration of high bitrate quantum random number generation with coherent laser light, Appl. Phys. Lett. **98**, 231103 (2011).

[32] J.-Y. Haw, S. Assad, A. Lance, N. Ng, V. Sharma, P. K. Lam, and T. Symul, Maximization of extractable randomness in a quantum random-number generator, Phys. Rev. Applied **3**, 054004 (2015).

[33] D. G. Marangon, G. Vallone, and P. Villoresi, Source-device-independent ultrafast quantum random number generation, Phys. Rev. Lett. **118**, 060503 (2017).

[34] M. Avesani, D. G. Marangon, G. Vallone, and P. Villoresi, Source-device-independent heterodyne-based quantum random number generator at 17 gbps, Nat. Commun. **9**, 5365 (2018).

[35] F. Raffaelli, *Quantum random number generators in integrated photonics*, Ph.D. thesis, University of Bristol (2019).

[36] Z. Zheng, Y. Zhang, W. Huang, S. Yu, and H. Guo, 6 gbps real-time optical quantum random number generator based on vacuum fluctuation, Rev. Sci. Instrum. **90**, 043105 (2019).

[37] Q. Zhou, R. Valivarthi, C. John, and W. Tittel, Practical quantum random-number generation based on sampling vacuum fluctuations, Quantum Engineering **1**, e8 (2019).

[38] X. Guo, C. Cheng, M. Wu, Q. Gao, P. Li, and Y. Guo, Parallel real-time quantum random number generator, Opt. Lett. **44**, 5566 (2019).

[39] J.-R. Álvarez, S. Sarmiento, J. A. Lázaro, J. M. Gené, and J. P. Torres, Random number generation by coherent detection of quantum phase noise, Opt. Express **28**, 5538 (2020).

[40] D. Drahi, N. Walk, M. J. Hoban, A. K. Fedorov, R. Shakhovoy, A. Feimov, Y. Kurochkin, W. S. Kolthammer, J. Nunn, J. Barrett, *et al.*, Certified quantum random numbers from untrusted light, Phys. Rev. X **10**, 041048 (2020).

[41] P. R. Smith, D. G. Marangon, M. Lucamarini, Z. Yuan, and A. Shields, Simple source device-independent continuous-variable quantum random number generator, Phys. Rev. A **99**, 062326 (2019).

[42] A. Uchida, K. Amano, M. Inoue, K. Hirano, S. Naito, H. Someya, I. Oowada, T. Kurashige, M. Shiki, S. Yoshimori, *et al.*, Fast physical random bit generation with chaotic semiconductor lasers, Nat. Photonics **2**, 728 (2008).

[43] I. Kanter, Y. Aviad, I. Reidler, E. Cohen, and M. Rosenbluh, An optical ultrafast random bit generator, Nat. Photonics **4**, 58 (2010).

[44] C. Gabriel, C. Wittmann, D. Sych, R. Dong, W. Mauerer, U. L. Andersen, C. Marquardt, and G. Leuchs, A generator for unique quantum random numbers based on vacuum states, Nat. Photonics **4**, 711 (2010).

[45] Y. Shen, L. Tian, and H. Zou, Practical quantum random number generator based on measuring the shot noise of vacuum states, Phys. Rev. A **81**, 063814 (2010).

[46] X. Ma, F. Xu, H. Xu, X. Tan, B. Qi, and H.-K. Lo, Post-processing for quantum random-number generators: Entropy evaluation and randomness extraction, Phys. Rev. A **87**, 062327 (2013).

[47] Y. Shi, B. Chng, and C. Kurtsiefer, Random numbers from vacuum fluctuations, Appl. Phys. Lett. **109**, 041101 (2016).

[48] Y.-i. Hayashi, T. Sugawara, Y. Kayano, N. Homma, T. Mizuki, A. Satoh, T. Aoki, S. Minegishi, H. Sone, and H. Inoue, An analysis of information leakage from a cryptographic hardware via common-mode current, EMC'09 (2009).

[49] T. Steinle, J. N. Greiner, J. Wrachtrup, H. Giessen, and I. Gerhardt, Unbiased all-optical random-number generator, Phys. Rev. X **7**, 041050 (2017).

[50] M. Ramdani, E. Sicard, A. Boyer, S. B. Dhia, J. J. Whalen, T. H. Hubing, M. Coenen, and O. Wada, The electromagnetic compatibility of integrated circuits—past, present, and future, IEEE T. Electromagn. C. **51**, 78 (2009).

[51] I. Giechaskiel, Y. Zhang, and K. B. Rasmussen, A framework for evaluating security in the presence of signal injection attacks, in *European Symposium on Research in Computer Security* (Springer, 2019) pp. 512–532.

[52] A. Kerckhoffs, La cryptographic militaire, Journal des sciences militaires , 5 (1883).

[53] R. Hughes and J. Nordholt, Strengthening the security foundation of cryptography with whitewood's quantum-powered entropy engine (2016).

[54] W. A. Radasky, C. E. Baum, and M. W. Wik, Introduction to the special issue on high-power electromagnetics (hpem) and intentional electromagnetic interference (iemi), IEEE T. Electromagn. C. **46**, 314 (2004).

[55] Y. Zhang, K. Rasmussen, I. Giechaskiel, K. Rasmussen, J. Szefer, D. Antonioli, N. O. Tippenhauer, K. Rasmussen, D. Antonioli, N. O. Tippenhauer, *et al.*, Detection of electromagnetic interference attacks on sensor systems, in *IEEE Symposium on Security and Privacy*, Vol. 19 (ACM).

[56] M. Bucci and R. Luzzi, Design of testable random bit generators, in *International Workshop on Cryptographic Hardware and Embedded Systems* (Springer, 2005) pp. 147–156.

[57] V. Fischer, A closer look at security in random number generators design, in *International Workshop on Constructive Side-Channel Analysis and Secure Design* (Springer, 2012) pp. 167–182.

[58] H. Qin, R. Kumar, V. Makarov, and R. Alléaume, Homodyne-detector-blinding attack in continuous-

variable quantum key distribution, Phys. Rev. A **98**, 012312 (2018)