Reply to Reviewer comments to manuscript APP24-AR-01137

Dear Editor,

first, we would like to thank the reviewers for their careful reading of the manuscript, and for their positive remarks as well as their constructive suggestions. We try to address the points by the referees below:

Reviewer 1

The paper "Countering detector manipulation attacks in quantum communication through detector self-testing" presents an experimental demonstration of a new method for limiting attacks on single photon detectors related to quantum key distribution (QKD) protocol. QKD is the first commercial product of the second quantum revolution and although there are many companies around the world that are deploying QKD systems, there are many issues that have to be solved for closing implementation loopholes. Indeed, apart from the side-channel attacks, one important vulnerability of single-photon detectors is the blinding attacks. However, most of the countermeasures are expensive or require extra equipment which can limit the implementation. The authors here present a new method based on self-testing of detectors to confirm their intended operation without relying on specific aspects of their inner workings. In addition, the authors experimentally demonstrate this countermeasure with standard InGaAs avalanche photodetector. There are minor things (reported below) that have to be fixed, but in general, I would say that I am confident in suggesting the publication. General comments:

1. which difference do you expect if you use single-photon detectors in your experiment?

Answer:
We thank the reviewer for the valuable comments. Our countermeasure method can be applied to all single-photon detectors that could be manipulated by macroscopic light, thermal, or other mechanisms. This includes not only the InGaAs APD used for experimental demonstration in this paper, but also more sophisticated superconducting single-photon detectors. Our method does not alter the specific detector mechanism, nor makes specific assumptions about its physical nature; instead, it introduces the concept of a self-test to verify whether the detector remains sensitive to single photons.

We added a statement in the last paragraph of the conclusion to emphasize this important fact.

2. What about using different wavelengths from the eavesdropper?

Answer:
While we applied a band-pass filter in the receiver to block unwanted wavelengths (as it is

likely customary in any QKD receiver), the single-photon response of the detector should be similar across different wavelengths for the test light source and the expected signal wavelength. Even if we did not use a band-pass filter and the detector were blinded by a different wavelength, the overall scheme would still be able to successfully identify the attack. This is because all our countermeasure schemes rely on the fact that a blinded detector cannot respond to single photons. Therefore, it is not necessary to model the detector's response to different wavelengths in detail.

We did not add a specific remark on this to the manuscript, as we feel it would probably only distract from the main idea of the self-testing concept.

3. Do you need to recalibrate the device? And if yes how often?

Answer:
Similar to the answer to the first question, our method does not alter the specific detector mechanism, or relies on careful measurements of auxiliary detector parameters, like a reverse bias voltage of an APD. Our method does thus not require a careful calibration, including detection efficiency for detecting manipulations, as manipulated and non-manipulated detector event statistics are very different, and depend not strongly on uncertainties in the self-testing power. We believe this is a key advantage of our approach compared to other methods.

We thank the reviewer for pointing this out, and have included a corresponding sentence at the end of the first paragraph of the conclusion in the manuscript: "Contrary to efficiency variation and monitoring mechanisms to detect single photon detector manipulations, this scheme does not require a careful calibration, as manipulated and non-manipulated detector event statistics under self-testing are very different, and depend not strongly on uncertainties in the self-testing power."

4. I guess the problem is during the dead time of the detector. Can Eve exploit this dark zone?

Answer:
In the third example, self-blinding could introduce a long dead time, which could be exploited by Eve. To mitigate this, a delay comparable to the self-blinding time could be added before announcing the result in the classical channel. If the reviewer refers to the dead time caused by single-photon detection, during this period, Eve could send strong pulses to create fake states. These fake states could be identified by performing a second-order correlation on detection events from the same detector, as part of a typical statistics monitoring in a practical QKD implementation. Our third example of self-blinding can also reliably detect the attack by monitoring unexpected detection events during the self-blinding period.

However, our second example suffers a very low impact of the effective detector efficiency reduction. As we go through specific numbers of detection probabilities in a paragraph after equation 2, we expanded the estimation there in the manuscript. Please also refer to our response to the question 1 of referee 2.

5. I think it is important to stress that the new method can be employed with all the

single-photon detectors.

Answer:
We appreciate the reviewer for highlighting this and have incorporated it into the final paragraph of the paper.

Reviewer 2

In this manuscript, Shen and Kurtsiefer describe a scheme to detect detector blinding attacks in a QKD scenario. It is a based on including, at random times, a test signal from an additional local light source, the presence (or absence) of measurements thereof gives a statistical signature of legitimacy (or blinding attempts) during a QKD protocol. It is assumed that the additional light source is under full local control of the communicating party. Essentially, the random but controlled light source can be used to manipulate the background noise or single-photon sensitivity of a detector in a controlled way, which cannot be spoofed by the detector blinding attack. While I am sure that technical loopholes cannot be ruled out, the approach certainly seems to be offer very high specificity in identifying attacks. A full information theoretic analysis is not included, but would go beyond the scope of the work as a proof-of-principle demonstration.

1. One aspect which is not completely clear to me is the trade-off between down-time when a blinding self-test takes place, and the up-time (data rate) of the protocol. If I understand correctly, this method is not a passive observer of detector blinding attacks, since measurement results of the QKD protocol obtained while "self-testing" cannot contribute to the QKD bit rate. Therefore I would be interested to see more details on the probability of identifying a blinding attack (presumably testing rate * success probability) versus QKD bit rate?

Answer:
We acknowledge the reviewer's concern regarding the trade-off between downtime and uptime. The reviewer is correct in noting that during a self-blinding attack, no secure keys can be transmitted. However, the time required to identify a blinding attack is very short.

As we go through explicit numbers estimating probabilities for discovering manipulations for our second example, we expanded this part (after equation 2) to estimate realistic detection down times for what we believe are realistic testing conditions: "The attack detection probabilities exemplified here can be reached with a sparse testing density: assuming a realistic detector dead time of $\tau_D = 1\,\mu\mathrm{s}$ after a "true" single photon (or background) detection event, and a randomized self-test pulse rate of $r_t = 2000\,\mathrm{s}^{-1}$, the above probability P_S of confirming a non-manipulated detector can be reached within $T = n/r_t = 5\,\mathrm{ms}$, while the detector is not available for detection of signal photons for a fraction of $\eta_t = \tau_D r_t = 0.2\%$. Such a reduction of the useful signal detection rate due to self-testing is likely lower than the uncertainties due to other environmental factors in practical systems."

A similar analysis can be performed for the other examples, but in our view, this should be part of a security assessment of a specific implementation in a given device. We feel this would go beyond the presentation of the key idea in this manuscript.

2. Indeed, the work is motivated by comparing other countermeasures to blinding attacks, such as multiplexing, watchdog detectors, monitoring detector performancee etc, or even MDI-QKD, but I'm not yet convinced the operational complexity and drop in key rate of the present scheme offers significant advantages over other approaches. Some quantitative comparison would be nice, if that's possible.

Answer:
The single-photon detector is typically the most expensive component in QKD systems. Moreover, multiplexing detectors require additional electronics, which increases both cost and complexity of the system. It also requires no additional data acquisition resources. Therefore, compared to multiplexing detectors, the cost of our approach appears to be significantly smaller, and can be implemented e.g. with a conventional LED.

The minimum blinding power is on the order of 100 picowatts, while watchdog detectors should be sufficient to detect very weak blinding signals on the order of picowatts. If the attack is below the sensitivity threshold of the watchdog detector, it may go undetected. Depending on how the watchdog detector is implemented, a careful threshold must be set to identify the attack; otherwise, issues with false positives or false negatives could arise. Furthermore, an optical beam splitter typically induces a loss of more than 1% of the QKD signal. Again, such a mechanism also requires additional data acquisition resources for the watchdog.

Monitoring detector performance or varying the detector's efficiency at random intervals significantly increases the complexity of the detector circuitry that has to be implemented in the initial detector design. In a retrofit scenario, this is likely hard compared to adding the self-test light source for our self-testing suggestion.

Experimental MDI-QKD requires careful inference of single photons (or weak coherent pulses) at a remote location, which is both technically challenging and costly. This is one reason why most discrete-variable commercial QKD systems still rely on BB84- or BBM92-family protocols.

In summary, our method does not alter the specific detector mechanism. Instead, it introduces the concept of a self-test to verify whether the detector remains sensitive to single photons. As a result, we do not need to carefully calibrate the detector's performance. Additionally, we use low-cost hardware that can be easily retrofitted into existing QKD systems with minimal impact on the key rate.

3. Nevertheless, the paper is clear and very well written, with convincing results. I am sure this paper is of interest to the practical QKD community, however I am not convinced it is of sufficiently broad interest to the wider photonics research community outside of this field. Perhaps a more specialized journal would be more appropriate.

Answer:
We thank the reviewer for his positive view of our scheme. However, we believe that the

scheme we present does not only address a key vulnerability of discrete variable QKD systems that should be of interest to a technical QKD community in a simple way, but offers a concept of testing a hypothesis of proper single photon detection in a device-independent spirit, as we do not make any assumptions on the detection mechanism at all. We believe that such a test concept could be inspiring to a wider photonics community, and therefore feel that APL photonics would be a suitable platform for this work.

Reviewer 3

In this work, Shen and Kurtsiefer proposed three interesting countermeasures based on self-testing of single-photon detectors to address manipulation attacks by eavesdroppers in secure quantum communication applications. Specifically, they present a self-testing strategy capable of detecting both suppression and false event generation attacks with high statistical significance. This method requires only a simple light source near the detector and does not rely on assumptions about the nature of potential manipulation, effectively addressing even unknown attacks such as detector blinding. The proposed approach provides a practical and straightforward solution to mitigate hardware vulnerabilities in quantum key distribution (QKD) systems, making it easier to implement and retrofit compared to device-independent methods.

After carefully reviewing the manuscript, I find the work to be both interesting and impactful for secure quantum communication applications. The manuscript is well-written, well-organized, and easy to follow without significant technical difficulty. The topic aligns closely with the journal's scope, and the quality meets its high standards. I did not identify any significant technical errors in the paper. Therefore, I am happy to recommend its publication in APL Photonics.

Answer:
We greatly appreciate the thorough review and the kind recommendation for publication in APL Photonics.

With this, we hope to have addressed the points highlighted by the referees, and look forward for your reply.

We attach a difference file between the previous and amended manuscript for easy reference.

With Best Regards,

Lijiong Shen and Christian Kurtsiefer