Peer Review File

Manuscript Title: Experimental relativistic zero-knowledge proofs

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

The manuscript "Experimental relativistic zero knowledge proofs" experimentally demonstrates the zero-knowledge proof of the three color ability problem. Based on the theory "Practical Relativistic Zero-Knowledge for NP", the experiment adapts the multi-provers method and realized the ZKP with relativistic constraint, presenting a zero-knowledge proof without using a one-way function. It is an indeed interesting experiment.

However, my main concern is its significance. In the implementation, the two verifiers and two provers accomplish the mission with specially designed hardware and space-like separation between them. However, in the widely used classical zero-knowledge proofs, any users may adapt the zero-knowledge proof protocol anonymously. That is why zero-knowledge is taken as a primitive and core protocol in current ICT. As a comparison, there are stringent requirements of physical relations between the verifiers and provers in this protocol. The requirement of the space-like relation is quite difficult to meet for users. Actually, the requirement that verifiers and provers have to meet in person is in violation of the anonymous privacy that is implicitly required in the zero-knowledge proofs.

The work presented in this manuscript is solid in physics and sound in information theory. But the relativistic requirement makes it only attractive academically and limits the potential application in practice. Therefore, I wouldn’t suggest its publication of Nature due to the limited scalability and potential usage. It can be considered to be published in Nature Communication as long as the authors address my following minor comments.

Below are a few small comments:
1. The experiment demonstrates a graph with V~5000 and E~10000, the limit of the size of the graph should be considered.
2. The prover may have quantum entanglement and quantum memory as a resource. The security that the non-classical but possible correlation exists should be further considered.

Referee #2 (Remarks to the Author):

Referee Report on *Experimental relativistic zero-knowledge proofs*

Nature manuscript 2021-01-00724

referee report dated 25 March 2021

Having read this manuscript (including the Methods supplement) with great interest, I have no hesitation to recommend publication in *Nature* in the strongest possible terms. This is truly a groundbreaking experimental proof that what may have seemed to be unrealistic theory at best can actually be implemented and might even become practical in the future. Furthermore, the paper is written in clear terms that are easy to follow without requiring specialized training. The motivations are clearly stated and convincing. I am confident that this paper will appeal to a broad class of readers.

The notion of *zero-knowledge proofs* is of such recognized importance that Shafi Goldwasser, Silvio Micali and Charles Rackoff were awarded the very first *Gödel Prize* [1] in 1993 for their groundbreaking paper "The knowledge complexity of interactive proof systems" [1], which introduced the notion. Simply said, a zero-knowledge proof for a statement would convince a "verifier" that the statement is true, but without giving any hint on how to actually prove it. This turns out to have tremendous applications in cryptographic protocols, for example for identification purposes. A few years later, Avi Wigderson (who just won the prestigious Abel Prize) joined forces with Micali and Oded Goldreich to prove that all statements in NP (whose proofs can be verified efficiently) can be proved in zero-knowledge [3] (this was also discovered independently by Claude Crépeau, an author of the current paper, with Gilles Brassard). General zero-knowledge proofs hinge upon the notion of *bit commitment*. However, it seemed initially that bit commitments (and therefore zero-knowledge proofs) could at best be secure from a computational perspective, meaning that at least one of the parties could cheat given sufficient computing power, and assumptions were needed even to infer computational security. This is why the subsequent [2] discovery that unconditionally secure bit commitment is possible in the context of multiple provers [9] is so important.

[1] The Gödel Prize is "an annual prize for outstanding papers in the area of theoretical computer science".

[2] The chronology of these discoveries is muddled terribly by the year of their publication due to unequal delay between the original discoveries and publication dates, and the fact that some papers appeared only in conference proceedings whereas others appeared (after significant delay) in archival journals long after they were first published in conference proceedings. For instance, Ref. [9], dated 1988 in the bibliography of the paper being refereed, is in fact posterior to Refs. [1, 3], dated 1989 and 1991. In this case, the reason is that Ref. [9] never appeared in a journal, contrary to Refs. [1, 3], and 1988 is the year of the conference at which it was presented, which was subsequent to the conferences at which Refs. [1, 3] were presented.

Despite the beautiful theory of multi-prover schemes, their security can only be established unconditionally if the provers are unable to communicate between them fast enough to cheat by making one's response to a challenge depend on the challenge asked to the other prover. In turn, this can only be enforced by exquisite synchronization, very fast response time, and sufficient distance, owing to the nonsignalling principle[3] of Einstein's special relativity. This is why the current paper is about *relativistic* zero-knowledge proofs. There is nothing new in what I wrote so far since the idea of basing cryptographic security on the theory of relativity is a thirty year old idea due to Jo Kilian [10] and made popular by Adrian Kent [11].

What makes the current paper groundbreaking and definitely worthy of *Nature* publication is that, to the best of my knowledge, this idea had never been implemented experimentally to demonstrate a nontrivial statement in zero-knowledge. The reason for this is that all known explicit protocols to exploit those theoretical ideas would require an unmanageable amount of memory and communication, an unreasonable distance between the provers, and/or unfeasibly quick responses to complicated challenges. The game changer was a protocol proposed last year [12] by one of the authors of the current paper (Claude Crépeau). For the first time, it became reasonable to hope for an implementation of the dream that arose three decades ago. And this is precisely what this paper offers, or rather it reports on *two* different implementations of these ideas. Note that even knowing the ideas behind Ref. [12], much still needed to be fine tuned, both experimentally and theoretically, before this quest could become reality. Although this remains at the level of a proof of existence, and the current proofs of unconditional security do not cover the case of quantum cheaters, the door is now open to potential practical applications.

Obviously, my enthusiastic recommendation is to accept this paper for publication in *Nature*. Nevertheless, being a conscientious referee, I read the paper with exacting precision. And naturally, I found a few problems that need fixing. But except for taking care of my more important remarks, denoted below by asterisks ("$\star$"), this paper could go to print even if none of my suggestions were taken into consideration.

In what follows, I refer to the page number ("p"), the side ("L" for left of "R" for right) and the line number ("ℓ") according to the submitted manuscript dated 19 December 2020. The absolute value of negative line numbers should be counted up from the bottom of the page. The more important remarks are identified with an asterisk ("$\star$"). In the case of page 1, the line numbers start after the Abstract (about which I have no comments). Similarly, line numbers start after the caption in pages that start with a double-column figure. However, comments on captions are denoted by page ("p") and line ("ℓ") numbers *without* the "L" or "R" modifier. I won't apologize if this is confusing because YOU (the *Nature* editors) are not making my task easy by not providing line numbers and not allowing me to mark my corrections directly in the submitted pdf!

[3] The *nonsignalling principle* asserts that information cannot be transmitted faster than at the speed of light.

Specific Comments

– $p1\mathsf{L}\ell14$. For someone who is not aware of the contents of my second footnote (on page 1 of this report), referring to a paper [1] published in 1989 seems incoherent with the (correct) statement that it initiated a field that “was born in the *middle* of the 1980’s”!

– $p1\mathsf{R}\ell2$. A reference to Cook’s seminal paper and/or the influential book of Garey and Johnson might be appropriate references on the first mention of the class NP.

– $p2$, FIG. 1. I don’t think you explain why the two up-arrows go only 4/5 the distance and then are dotted after the arrowhead. I have my guess about the reason, but I should not have had to guess!

– $p2\ell4$. I could see no reason for having the yellow nodes *dashed* and the red nodes *dotted*. Why should the colours be insuffficient? Ah, perhaps it is in case the page is not printed in colour, or perhaps to accommodate colour-blind readers?

– $p2\ell7$. I suggest replacing “special relativity” by “the nonsignalling principle of special relativity”.

⋆ $p2\ell13$. Statement “the consistency of the provers’ answers can then be tested” is problematic because I don’t see how this can be done with the protocol presented here. In fact, the protocol that was implemented is the one from Ref. [12], which is explained in the Methods supplement, and makes it possible indeed to test the provers’ answers for consistency, again as explained in the Methods supplement. Note that it is admitted on $p3\mathsf{R}\ell28$ that the protocol described in Figure 1 is “a pedagogical variant” of the protocol that was actually implemented. Is it possible to “test the consistency of the provers’ answers” with this pedagogical variant? If yes, why did you implement something slightly more complicated? If no, you should at least say upfront that this pedagogical variant is given for simplicity but that it does not actually fit the bill.

– $p3\mathsf{L}\ell\ell10$–12. You say that you “discuss the prospects of extending the security to the general case of quantum-mechanically correlated provers”, which you do indeed. However, you do not mention the possibility of extending the security *proof* of the protocol as implemented to cover the case of would-be cheating quantum provers. I realize that the current proof techniques would require the implementation of unreasonable protocols to cover this case, but could it not be that your implementation is quantum-secure as it is? If I had to guess, I would bet that it is! I suggest you mention that this is an open problem.

⋆ $p3\mathsf{L}\ell13$. You should say up-front that the protocol described here and below is “a pedagogical variant” of the protocol that was actually implemented, rather than postponing this admission to $p3\mathsf{R}\ell28$.

– $p3\mathsf{L}\ell\ell 30$–31. It is not clear at this point that the (i, j, b) provided by each verifier to their prover can be different, i.e. that one verifier provides (i, j, b) to its prover whereas the other can provide a possibly different (i', j', b'). This becomes clear fairly quickly below, but at this point in reading the paper it is natural to incorrectly assume that the two challenges are identical because they are both called (i, j, b).

– $p3\mathsf{L}\ell\ell 37$–39,43–44. The repetition of the fact that "the provers know that the graph is three-colourable" seems redundant to me.

– $p3\mathsf{L}\ell{-8}$. I would replace the word "ignore" by "take no account of" (just a suggestion).

– $p3\mathsf{R}\ell 28$. It's about time you admit that we've been discussing "a pedagogical variant" of the protocol that was actually implemented!

– $p3\mathsf{R}\ell 34$. Perhaps the title of this section should be "The graph" rather than "Graphs" since only one graph was used (again, just a suggestion).

– $p3\mathsf{R}\ell\ell 34$–42. This entire paragraph is very redundant with $p1\mathsf{R}\ell\ell 1$–10. It may be OK to repeat this material, but if you are short in space (for example to explain properly that the protocol given here is but a pedagogical variant of the one that was really implemented), this may be a good place to cut.

– $p3\mathsf{R}\ell 46$. The words "algorithms efficient" should be swapped: "efficient algorithms".

– $p3\mathsf{R}\ell\ell 55$–56. You may wish to avoid this repetition of "so that".

⋆ $p4\mathsf{L}\ell{-19}$. I am a little worried by the fact that the verifiers do not actively control the actual position of the provers in space. The verifiers trust each other and therefore they know that they are separated by 390 m. But is it clear that the provers could not cheat if they were closer to one another than the verifiers think they are? Should you not have used *distance bounding* for the verifiers to check that the provers are at some maximum distance δ from them, and therefore that the provers are at least $390\,\mathrm{m} - 2\delta$ apart from each other. If this is not necessary, perhaps you should explain why.

– $p4\mathsf{R}\ell\ell 10$–11. Rather than saying that "A trigger signal is exchanged between the two verifiers", it would be more accurate to write that "A trigger signal is sent from the first verifier to the second".

– $p4\mathsf{R}\ell{-13}$. One metre apart could suffice… WOW! Fabulous! Indeed my hands are more than one metre apart when I outstretch my arms.

– $p5\mathsf{L}\ell 5$. Perhaps add "which is" in front of "completely". Also, perhaps say explicitly that this would require more than 3×10^{15} rounds!

– $p6\mathsf{L}\ell 7$. I see that there are two figures number 2: one above and one below the caption. Nevertheless, I find the expression "Figs 2" strange because of the plural Figs followed by the singular number 2.

– $p6\mathsf{L}\ell10$. It was here, when I saw $a_1 \equiv b_i \cdot r + c_i$, that I understood that the protocol described in the main text is not the one that was actually used in the experiment. (I guess I had missed the admission at $p3\mathsf{R}\ell28$ that the protocol described in the main text is "a pedagogical variant" of the protocol used in the experiment.) It may be useful to start this section 1 of the Methods supplement by stating explicitly that we are now going to describe the actually implemented protocol and repeating that it is slightly different from the one described in the main text.

– $p6\mathsf{L}\ell23$. Hmmm... Here we have that $a_1 + a'_1 = [b_i r + c_i] + [b_i(-r) + c_i] \equiv 2c_i \pmod 3$ and similarly $a_2 + a'_2 \equiv 2c_j \pmod 3$. It follows that it is correct to conclude that $c_i \neq c_j$ if (and only if) $a_1 + a'_1 \not\equiv a_2 + a'_2 \pmod 3$ as claimed, but only because $x = y \pmod 3 \Leftrightarrow 2x = 2y \pmod 3$. I am mentioning this because the main text may make the careless reader believe erroneously that $a_1 + a'_1 = c_i$ and $a_2 + a'_2 = c_j$.

– $p6\mathsf{R}\ell7$. Again, I would replace the word "ignore" by "take no account of".

⋆ $p6\mathsf{R}\ell{-20}$. I strongly object to writing "In Ref. [13] the authors give [...]" rather than naming said authors explicitly, especially in this case when there are only two authors.

– $p8\mathsf{L}\ell27$. I would replace the word "electronical" with "electronic". When I checked for "electronical" with the help of Google, I was redirected to the *Urban Dictionary*, which says that this word is "a drunken reference to electronic equipment". As for the word "electronicals" (which does not appear in the paper, but I could not resist the fun of mentioning it), it would be "anything using a power source not understood by the user"! `:-)`

– $p8\mathsf{L}\ell32$. This is the only section that I did not try to understand in detail. However, is it obvious that the zero-knowledge property of the protocol can still be formally proven with the simulator definition?

– $p8\mathsf{L}\ell34$–35. This is slightly ambiguous. You mean that in each round both provers need to use the same random information, **not** that the same random information is used round after round!

Referee #3 (Remarks to the Author):

The idea of zero-knowledge (ZK) proof was introduced in theoretical cryptography in the 1980s. Since then it has found many applications, including in applied cryptography such as to identification protocols.

It is known that secure ZK proofs can be constructed under computational assumptions such as the existence of one-way functions. It is also possible to construct ZK proofs that are secure based on an assumption of spatial isolation between two provers. This has been known since the 1980s as well. From a theoretical point of view the advantage of such proof systems is that one does not need to rely on the hardness of a computational problem for the protocol to be secure. The inconvenient is that in practice spatial isolation is enforced using relativistic separation, which at reasonable distances requires an extremely fast interaction.

This paper reports on an experimental demonstration of a well-known ZK proof of the second kind, the so-called GMW protocol from 1991 (ref. [3]). The experiment involves setting up two distant communication links so that a round-trip communication can be performed at each distant space faster than the time it takes for light to travel between the two locations. As long as this can be ensured then the protocol is guaranteed to be sound: security rests on a "relativistic", as opposed to "computational", assumption.

From the theoretical point of view the theory used in the paper has been well-known since the early 1990s. From a practical point of view it is not clear if there is any real use for relativistic ZK protocols, because computational protocols are widely used and I don't think there is any problem with them (e.g. the assumption of one-way functions is considered safe). Of course one could arguably say the same of quantum key distribution, so it is not a very fair criticism. For me the main interest of the paper lies in a "proof of principle" showing that protocols that may a priori seem completely out of reach can in fact be implemented using pretty straightforward equipment. An important drawback is that the protocol is not known to be secure in case the malicious parties may make use of quantum entanglement to cheat.

I would like to know how the experimental setup differs from, or improves on, the one reported in [15]. It is known that ZK proofs can be constructed from secure bit commitments, so could the protocol from [15] not be adapted in a straightforward manner to obtain siilar results as reported here?

Additional comments:

* In the first page, it would be worth explaining the notion of a one way function when first introducing it, as this may not be part of the basic knowledge of a Nature reader.

* p.3 1st column: "pre-agree on random three-colourings", the colourings should be proper.

* Your protocol makes use of two verifiers and two provers. This is somewhat unexpected, as usually ZK proofs involve a single verifier and two provers. Could you explain why? Does the use of two verifiers make the experiment more accessible?

Referee Report on *Experimental relativistic zero-knowledge proofs*

Nature manuscript 2021-01-00724

referee report dated 25 March 2021

Having read this manuscript (including the Methods supplement) with great interest, I have no hesitation to recommend publication in *Nature* in the strongest possible terms. This is truly a groundbreaking experimental proof that what may have seemed to be unrealistic theory at best can actually be implemented and might even become practical in the future. Furthermore, the paper is written in clear terms that are easy to follow without requiring specialized training. The motivations are clearly stated and convincing. I am confident that this paper will appeal to a broad class of readers.

The notion of *zero-knowledge proofs* is of such recognized importance that Shafi Goldwasser, Silvio Micali and Charles Rackoff were awarded the very first *Gödel Prize* [1] in 1993 for their groundbreaking paper "The knowledge complexity of interactive proof systems" [1], which introduced the notion. Simply said, a zero-knowledge proof for a statement would convince a "verifier" that the statement is true, but without giving any hint on how to actually prove it. This turns out to have tremendous applications in cryptographic protocols, for example for identification purposes. A few years later, Avi Wigderson (who just won the prestigious Abel Prize) joined forces with Micali and Oded Goldreich to prove that all statements in NP (whose proofs can be verified efficiently) can be proved in zero-knowledge [3] (this was also discovered independently by Claude Crépeau, an author of the current paper, with Gilles Brassard). General zero-knowledge proofs hinge upon the notion of *bit commitment*. However, it seemed initially that bit commitments (and therefore zero-knowledge proofs) could at best be secure from a computational perspective, meaning that at least one of the parties could cheat given sufficient computing power, and assumptions were needed even to infer computational security. This is why the subsequent [2] discovery that unconditionally secure bit commitment is possible in the context of multiple provers [9] is so important.

[1] The Gödel Prize is "an annual prize for outstanding papers in the area of theoretical computer science".

[2] The chronology of these discoveries is muddled terribly by the year of their publication due to unequal delay between the original discoveries and publication dates, and the fact that some papers appeared only in conference proceedings whereas others appeared (after significant delay) in archival journals long after they were first published in conference proceedings. For instance, Ref. [9], dated 1988 in the bibliography of the paper being refereed, is in fact posterior to Refs. [1, 3], dated 1989 and 1991. In this case, the reason is that Ref. [9] never appeared in a journal, contrary to Refs. [1, 3], and 1988 is the year of the conference at which it was presented, which was subsequent to the conferences at which Refs. [1, 3] were presented.

Despite the beautiful theory of multi-prover schemes, their security can only be established unconditionally if the provers are unable to communicate between them fast enough to cheat by making one's response to a challenge depend on the challenge asked to the other prover. In turn, this can only be enforced by exquisite synchronization, very fast response time, and sufficient distance, owing to the nonsignalling principle[3] of Einstein's special relativity. This is why the current paper is about *relativistic* zero-knowledge proofs. There is nothing new in what I wrote so far since the idea of basing cryptographic security on the theory of relativity is a thirty year old idea due to Jo Kilian [10] and made popular by Adrian Kent [11].

What makes the current paper groundbreaking and definitely worthy of *Nature* publication is that, to the best of my knowledge, this idea had never been implemented experimentally to demonstrate a nontrivial statement in zero-knowledge. The reason for this is that all known explicit protocols to exploit those theoretical ideas would require an unmanageable amount of memory and communication, an unreasonable distance between the provers, and/or unfeasibly quick responses to complicated challenges. The game changer was a protocol proposed last year [12] by one of the authors of the current paper (Claude Crépeau). For the first time, it became reasonable to hope for an implementation of the dream that arose three decades ago. And this is precisely what this paper offers, or rather it reports on *two* different implementations of these ideas. Note that even knowing the ideas behind Ref. [12], much still needed to be fine tuned, both experimentally and theoretically, before this quest could become reality. Although this remains at the level of a proof of existence, and the current proofs of unconditional security do not cover the case of quantum cheaters, the door is now open to potential practical applications.

Obviously, my enthusiastic recommendation is to accept this paper for publication in *Nature*. Nevertheless, being a conscientious referee, I read the paper with exacting precision. And naturally, I found a few problems that need fixing. But except for taking care of my more important remarks, denoted below by asterisks ("$\star$"), this paper could go to print even if none of my suggestions were taken into consideration.

In what follows, I refer to the page number ("p"), the side ("L" for left of "R" for right) and the line number ("ℓ") according to the submitted manuscript dated 19 December 2020. The absolute value of negative line numbers should be counted up from the bottom of the page. The more important remarks are identified with an asterisk ("$\star$"). In the case of page 1, the line numbers start after the Abstract (about which I have no comments). Similarly, line numbers start after the caption in pages that start with a double-column figure. However, comments on captions are denoted by page ("p") and line ("ℓ") numbers *without* the "L" or "R" modifier. I won't apologize if this is confusing because YOU (the *Nature* editors) are not making my task easy by not providing line numbers and not allowing me to mark my corrections directly in the submitted pdf!

[3] The *nonsignalling principle* asserts that information cannot be transmitted faster than at the speed of light.

Specific Comments

– $p1\mathsf{L}\ell14$. For someone who is not aware of the contents of my second footnote (on page 1 of this report), referring to a paper [1] published in 1989 seems incoherent with the (correct) statement that it initiated a field that "was born in the *middle* of the 1980's"!

– $p1\mathsf{R}\ell2$. A reference to Cook's seminal paper and/or the influential book of Garey and Johnson might be appropriate references on the first mention of the class NP.

– $p2$, FIG. 1. I don't think you explain why the two up-arrows go only 4/5 the distance and then are dotted after the arrowhead. I have my guess about the reason, but I should not have had to guess!

– $p2\ell4$. I could see no reason for having the yellow nodes *dashed* and the red nodes *dotted*. Why should the colours be insuffficient? Ah, perhaps it is in case the page is not printed in colour, or perhaps to accommodate colour-blind readers?

– $p2\ell7$. I suggest replacing "special relativity" by "the nonsignalling principle of special relativity".

⋆ $p2\ell13$. Statement "the consistency of the provers' answers can then be tested" is problematic because I don't see how this can be done with the protocol presented here. In fact, the protocol that was implemented is the one from Ref. [12], which is explained in the Methods supplement, and makes it possible indeed to test the provers' answers for consistency, again as explained in the Methods supplement. Note that it is admitted on $p3\mathsf{R}\ell28$ that the protocol described in Figure 1 is "a pedagogical variant" of the protocol that was actually implemented. Is it possible to "test the consistency of the provers' answers" with this pedagogical variant? If yes, why did you implement something slightly more complicated? If no, you should at least say upfront that this pedagogical variant is given for simplicity but that it does not actually fit the bill.

– $p3\mathsf{L}\ell\ell10$–12. You say that you "discuss the prospects of extending the security to the general case of quantum-mechanically correlated provers", which you do indeed. However, you do not mention the possibility of extending the security *proof* of the protocol as implemented to cover the case of would-be cheating quantum provers. I realize that the current proof techniques would require the implementation of unreasonable protocols to cover this case, but could it not be that your implementation is quantum-secure as it is? If I had to guess, I would bet that it is! I suggest you mention that this is an open problem.

⋆ $p3\mathsf{L}\ell13$. You should say up-front that the protocol described here and below is "a pedagogical variant" of the protocol that was actually implemented, rather than postponing this admission to $p3\mathsf{R}\ell28$.

– $p3\mathsf{L}\ell\ell30$–31. It is not clear at this point that the (i, j, b) provided by each verifier to their prover can be different, i.e. that one verifier provides (i, j, b) to its prover whereas the other can provide a possibly different (i', j', b'). This becomes clear fairly quickly below, but at this point in reading the paper it is natural to incorrectly assume that the two challenges are identical because they are both called (i, j, b).

– $p3\mathsf{L}\ell\ell37$–39,43–44. The repetition of the fact that "the provers know that the graph is three-colourable" seems redundant to me.

– $p3\mathsf{L}\ell-8$. I would replace the word "ignore" by "take no account of" (just a suggestion).

– $p3\mathsf{R}\ell28$. It's about time you admit that we've been discussing "a pedagogical variant" of the protocol that was actually implemented!

– $p3\mathsf{R}\ell34$. Perhaps the title of this section should be "The graph" rather than "Graphs" since only one graph was used (again, just a suggestion).

– $p3\mathsf{R}\ell\ell34$–42. This entire paragraph is very redundant with $p1\mathsf{R}\ell\ell1$–10. It may be OK to repeat this material, but if you are short in space (for example to explain properly that the protocol given here is but a pedagogical variant of the one that was really implemented), this may be a good place to cut.

– $p3\mathsf{R}\ell46$. The words "algorithms efficient" should be swapped: "efficient algorithms".

– $p3\mathsf{R}\ell\ell55$–56. You may wish to avoid this repetition of "so that".

⋆ $p4\mathsf{L}\ell-19$. I am a little worried by the fact that the verifiers do not actively control the actual position of the provers in space. The verifiers trust each other and therefore they know that they are separated by 390 m. But is it clear that the provers could not cheat if they were closer to one another than the verifiers think they are? Should you not have used *distance bounding* for the verifiers to check that the provers are at some maximum distance δ from them, and therefore that the provers are at least $390\,\mathrm{m} - 2\delta$ apart from each other. If this is not necessary, perhaps you should explain why.

– $p4\mathsf{R}\ell\ell10$–11. Rather than saying that "A trigger signal is exchanged between the two verifiers", it would be more accurate to write that "A trigger signal is sent from the first verifier to the second".

– $p4\mathsf{R}\ell-13$. One metre apart could suffice… WOW! Fabulous! Indeed my hands are more than one metre apart when I outstretch my arms.

– $p5\mathsf{L}\ell5$. Perhaps add "which is" in front of "completely". Also, perhaps say explicitly that this would require more than 3×10^{15} rounds!

– $p6\mathsf{L}\ell7$. I see that there are two figures number 2: one above and one below the caption. Nevertheless, I find the expression "Figs 2" strange because of the plural Figs followed by the singular number 2.

- $p6\mathsf{L}\ell10$. It was here, when I saw $a_1 \equiv b_i \cdot r + c_i$, that I understood that the protocol described in the main text is not the one that was actually used in the experiment. (I guess I had missed the admission at $p3\mathsf{R}\ell28$ that the protocol described in the main text is "a pedagogical variant" of the protocol used in the experiment.) It may be useful to start this section 1 of the Methods supplement by stating explicitly that we are now going to describe the actually implemented protocol and repeating that it is slightly different from the one described in the main text.

- $p6\mathsf{L}\ell23$. Hmmm... Here we have that $a_1 + a'_1 = [b_i r + c_i] + [b_i(-r) + c_i] \equiv 2c_i \pmod 3$ and similarly $a_2 + a'_2 \equiv 2c_j \pmod 3$. It follows that it is correct to conclude that $c_i \neq c_j$ if (and only if) $a_1 + a'_1 \not\equiv a_2 + a'_2 \pmod 3$ as claimed, but only because $x = y \pmod 3 \Leftrightarrow 2x = 2y \pmod 3$. I am mentioning this because the main text may make the careless reader believe erroneously that $a_1 + a'_1 = c_i$ and $a_2 + a'_2 = c_j$.

- $p6\mathsf{R}\ell7$. Again, I would replace the word "ignore" by "take no account of".

⋆ $p6\mathsf{R}\ell{-20}$. I strongly object to writing "In Ref. [13] the authors give [...]" rather than naming said authors explicitly, especially in this case when there are only two authors.

- $p8\mathsf{L}\ell27$. I would replace the word "electronical" with "electronic". When I checked for "electronical" with the help of Google, I was redirected to the *Urban Dictionary*, which says that this word is "a drunken reference to electronic equipment". As for the word "electronicals" (which does not appear in the paper, but I could not resist the fun of mentioning it), it would be "anything using a power source not understood by the user"! `:-)`

- $p8\mathsf{L}\ell32$. This is the only section that I did not try to understand in detail. However, is it obvious that the zero-knowledge property of the protocol can still be formally proven with the simulator definition?

- $p8\mathsf{L}\ell34$–35. This is slightly ambiguous. You mean that in each round both provers need to use the same random information, **not** that the same random information is used round after round!

Summary of changes, all marked in red in the manuscript:

- More focused and clarified presentation of the protocol. We now use only the simplified version, detailed in the Methods section (part 1). This lightens the main text which now meets the Nature format.
- New references [1] and [4] added following the comments of Referee 2.
- Improved discussion on the case of quantum provers. The paragraph in the main text is now shortened and clarified, while a more thorough discussion is given in the Methods section (part 5).
- Numerous minor changes to address the comments of all Referees.

Referee 1

We thank the Referee for their time and useful feedback. Below we provide point-by-point answers to each question or comment of the Referee.

The manuscript "Experimental relativistic zero knowledge proofs" experimentally demonstrates the zero-knowledge proof of the three color ability problem. Based on the theory "Practical Relativistic Zero-Knowledge for NP", the experiment adapts the multi-provers method and realized the ZKP with relativistic constraint, presenting a zero-knowledge proof without using a one-way function. It is an indeed interesting experiment.

However, my main concern is its significance. In the implementation, the two verifiers and two provers accomplish the mission with specially designed hardware and space-like separation between them. However, in the widely used classical zero-knowledge proofs, any users may adapt the zero-knowledge proof protocol anonymously. That is why zero-knowledge is taken as a primitive and core protocol in current ICT. As a comparison, there are stringent requirements of physical relations between the verifiers and provers in this protocol. The requirement of the space-like relation is quite difficult to meet for users. Actually, the requirement that verifiers and provers have to meet in person is in violation of the anonymous privacy that is implicitly required in the zero-knowledge proofs.

As the Referee points out, some current trend in blockchain application uses zero-knowledge proofs for anonymity. There is however no inherent requirement of zero-knowledge that connects it to anonymity: anonymity and zero-knowledge have a broad life on their own. We motivate our work with the more traditional application of user authentication. Although authentication may appear to contradict anonymity, it is indeed not the *opposite* of anonymity. In many situations, anonymity is achieved via the assignment of pseudonyms and proving ownership of a pseudonym can be accomplished using zero-knowledge. It is also a misconception to believe verifiers and provers must meet *in person*. Actual verifiers and provers are devices (a teller machine, a debit card, etc) provided by their owners. Anonymity of their owners is not at stake here.

Another question raised by the Referee concerns the ease-of-implementation of our protocol. Indeed, at first sight, the requirement of space-like separation appears as a hurdle for practical applications, in particular on short distances. Nevertheless, as we show in our work, relativistic zero-knowledge can be implemented efficiently on a distance of tens of metres, using only standard off-the-shelf equipment. We believe that prospects for implementations on even shorter distances (down to the metre scale) are very promising, as we discuss in more detail below.

The work presented in this manuscript is solid in physics and sound in information theory.

We receive this opinion from the Referee with great enthusiasm.

But the relativistic requirement makes it only attractive academically and limits the potential application in practice.

Here, we have to respectfully disagree with the Referee. On the contrary, we believe that our (arguably low-tech) experimental proof-of-concept demonstrates that, with some investment, the relativistic scenario can be made quite practical even though the speed of light is very large.

In the next paragraphs we describe precisely the gain obtained by using more expensive and dedicated equipment. We estimate that readily available equipment (using faster, hence more expensive, FPGA boards) would enable an implementation with a separation of only few metres, i.e., within a large room (see details below). Furthermore, performance could still be significantly improved using dedicated electronic chips, e.g., application-specific integrated circuit (ASIC), designed for this specific protocol. This would allow to shorten the running time by a few nanoseconds, hence allowing for an implementation on even shorter distance (possibly down to one metre). In fact, such systems would eventually also turn out to be cheaper to produce. Let us imagine that a large (e.g., smartphone) company would be willing to pay for the barrier to entry (i.e., the development cost), then the individual cost of a single chip would be drastically reduced by virtue of the economy of scale.

Let us now discuss in detail a faster implementation based on more expensive, but readily available, FPGA equipment. A test of the performances achievable on a higher-end FPGA chip (Kintex UltraScale+ xcku5p costing around 1300€, recall that the off-the-shelf FPGA chips used in our implementation cost around 100€ each) has been made using Xilinx Vivado software. The prover's function was synthesised, placed, and routed; a post-implementation timing simulation was performed. All values (i, j, b, a_1, a_2) were routed on input/output (IO) pins, which would operate low-voltage differential signaling (LVDS).

This implementation would use a clock rate of 500 MHz, and three cycles would be needed to compute the prover's answers without giving any timing violations. The total time between the validation of the input signal (i, j, b) and the output signals (a_1, a_2) would be around 8.8 ns, arising from the clock cycles together with the IO delays of the LVDS buffers.

A verifier using the same technology communicating with this prover would need a similar order of magnitude to capture the answers and compensate for the inaccuracy of the trigger. So we can safely claim that the whole exchange could be done in less than 20 ns, allowing to place both systems at a distance of 6 m, i.e., within a room.

Therefore, I wouldn't suggest its publication of Nature due to the limited scalability and potential usage. It can be considered to be published in Nature Communication as long as the authors address my following minor comments.

We hope that the clarifications given in the above replies will convince the Referee that our paper is suitable for Nature. We believe the multi-disciplinarity nature of our research will connect with a broad audience. The paper was written with great care to be readable by a vast scientific readership, and further improved based on the useful feedback of all Referees.

Below are a few small comments:

1. *The experiment demonstrates a graph with $V \sim 5000$ and $E \sim 10000$, the limit of the size of the graph should be considered.*

We agree with the Referee that our manuscript was not exhaustive regarding this point. Our experiment was conducted with a graph with $V = 588$ and $E = 1097$. Using the same hardware, we could reach graphs with $V \sim 5000$ and $E \sim 10000$, as stated in the manuscript (page 4). With parallel communication and an application-specific integrated circuit, the time needed for one round could be reduced to a few nanoseconds, thus allowing to run a couple of hundreds of millions of rounds in about a second, which is a reasonable time for applications. Keeping our level of security ($k = 100$), this corresponds to graphs with V and E of the order of 10^5. We note that state-of-the-art equipment (as mentioned above) features enough memory to handle this. We have now added a sentence in the main text (page 4).

2. *The prover may have quantum entanglement and quantum memory as a resource. The security that the non-classical but possible correlation exists should be further considered.*

The case of quantum provers is indeed of great interest; in the new version we have extended our discussion on this point. Here are three arguments to be optimistic:

1. There already exist two adaptations of our protocol that are provably secure against quantum provers. First, by adding a third verifier-prover pair, following the method proposed in Ref. [24] (previously Ref. [22]). Second, by considering larger graphs with a number of vertices and edges quadratically bigger than the current number of vertices, following Ref. [25] (previously Ref. [23]). Unfortunately, at this stage, the implementation of either of these adaptations is out of reach in practice, as the required number of rounds is way too large. Nevertheless, there is clearly room for improvement on the security bounds, which may significantly reduce the number of rounds. Moreover, it may also be possible to combine both of these adaptations to obtain a more efficient protocol. As these extensions were not covered in details in the previous version, we have now written a separate section (Methods 5) to discuss these in detail.

2. We share the point of view of Referee 2, namely, that our current protocol as is may actually be sound against quantum provers. This is supported by the fact that not a single graph that is not three-colourable (classically) has been found to be quantum three-colourable. Consult Ref. [CMN+06] below for definition of *quantum k-colourability* and existing examples.

3. In any case, if any change at all is required to guarantee security in the case of quantum provers, it should definitely not be of paradigmatic nature. Both solutions we currently consider (as in 1. above) do not modify the protocol per se, as they are forms of wrapping around our basic zero-knowledge protocol. Hopefully our work will motivate further theoretical investigations to find tighter bounds reducing the number of rounds within reasonable limits. A quadratic increase in the number of rounds instead of a fourth power would be quite manageable (see Methods 5).

Referee 2

We thank the Referee for their time and useful feedback. Below we provide point-by-point answers to each question or comment of the Referee.

Having read this manuscript (including the Methods supplement) with great interest, I have no hesitation to recommend publication in Nature in the strongest possible terms. This is truly a groundbreaking experimental proof that what may have seemed to be unrealistic theory at best can actually be implemented and might even become practical in the future. Furthermore, the paper is written in clear terms that are easy to follow without requiring specialized training. The motivations are clearly stated and convincing. I am confident that this paper will appeal to a broad class of readers. [. . .]

We thank the Referee for the positive comments.

Specific Comments

- *p1Lℓ14. For someone who is not aware of the contents of my second footnote (on page 1 of this report), referring to a paper [1] published in 1989 seems incoherent with the (correct) statement that it initiated a field that "was born in the middle of the 1980's"!*

This is not uncommon in computer science. There is indeed half a decade between the initial conference publication and the journal version. We have added the reference from 1985.

- *p2ℓ4. I could see no reason for having the yellow nodes dashed and the red nodes dotted. Why should the colours be insufficient? Ah, perhaps it is in case the page is not printed in colour, or perhaps to accommodate colour-blind readers?*

We mostly have in mind the possibility that readers print our manuscript in black and white. Having different contours seemed the simplest option to accommodate this.

* *p2ℓ13. Statement "the consistency of the provers' answers can then be tested" is problematic because I don't see how this can be done with the protocol presented here. In fact, the protocol that was implemented is the one from Ref. [12], which is explained in the Methods supplement, and makes it possible indeed to test the provers' answers for consistency, again as explained in the Methods supplement. Note that it is admitted on p3Rℓ28 that the protocol described in Figure 1 is "a pedagogical variant" of the protocol that was actually implemented. Is it possible to "test the consistency of the provers' answers" with this pedagogical variant? If yes, why did you implement something slightly more complicated? If no, you should at least say upfront that this pedagogical variant is given for simplicity but that it does not actually fit the bill.*

* *p3Lℓ13. You should say up-front that the protocol described here and below is "a pedagogical variant" of the protocol that was actually implemented, rather than postponing this admission to p3Rℓ28.*

- *p3Rℓ28. It's about time you admit that we've been discussing "a pedagogical variant" of the protocol that was actually implemented!*

- *p6Lℓ10. It was here, when I saw* $a_1 \equiv b_i \cdot r + c_i \pmod 3$*, that I understood that the protocol described in the main text is not the one that was actually used in the experiment. (I guess I had missed the admission at p3Rℓ28 that the protocol described in the main text is "a pedagogical variant" of the protocol used in the experiment.) It may be useful to start this section 1 of the Methods supplement by stating explicitly that we are now going to describe the actually implemented protocol and repeating that it is slightly different from the one described in the main text.*

The Referee raised here a subtle but indeed very important point. For historical reasons, we had first implemented the exact protocol of Ref. [14] (previously Ref. [12]) simply because the simplified version was not yet available. Only after this first experiment have we gotten the "pedagogical variant", which turned out to be as secure as its ancestor on top of being way easier to explain. The first version of our paper reflected this chronology, but we do agree with the Referee that this was confusing. That is why we have decided to mount the experiment again to adapt it to the new simplified version of the protocol so that we can now claim to have successfully implemented it. The manuscript's presentation is then considerably lighter. Note that, for length reason, the presentation of the protocol is now delayed to Methods 1; the reader should nonetheless be able to get all important features from the caption of Fig. 1.

- *p3Lℓℓ10-12. You say that you "discuss the prospects of extending the security to the general case of quantum-mechanically correlated provers", which you do indeed. However, you do not mention the possibility of extending the security proof of the protocol as implemented to cover the case of would-be cheating quantum provers. I realize that the current proof techniques would require the implementation of unreasonable protocols to cover this case, but could it not be that your implementation is quantum secure as it is? If I had to guess, I would bet that it is! I suggest you mention that this is an open problem.*

We do agree with the Referee and have added a sentence in this way.

* *p4Lℓ-19. I am a little worried by the fact that the verifiers do not actively control the actual position of the provers in space. The verifiers trust each other and therefore they know that they are separated by* 390 m*. But is it clear that the provers could not cheat if they were closer to one another than the verifiers think they are? Should you not have used distance bounding for the verifiers to check that the provers are at some maximum distance* δ *from them, and therefore that the provers are at least* 390 m - 2δ *apart from each other. If this is not necessary, perhaps you should explain why.*

Indeed an important fact is at stake here: do the verifiers need to locate the provers at all? The answer is simply no. All that is needed by the verifiers is to check that direct communication between each other at the speed of light would arrive later than the answers of their corresponding provers. If this is satisfied, the provers cannot possibly have communicated the same information between each other and reply faster.

- *p8Lℓ32. This is the only section that I did not try to understand in detail. However, is it obvious that the zero-knowledge property of the protocol can still be formally proven with the simulator definition?*

Absolutely. As a matter of fact even a stronger definition of zero-knowledge can be proven here: no-signalling simulators can simulate the conversation in polynomial time without getting back together. But this is another story altogether; consult Ref. [CY19] below.

- *p1Rℓ2. A reference to Cook's seminal paper and/or the influential book of Garey and Johnson might be appropriate references on the first mention of the class NP.*
- *p2, FIG. 1. I don't think you explain why the two up-arrows go only 4/5 the distance and then are dotted after the arrowhead. I have my guess about the reason, but I should not have had to guess!*
- *p2ℓ7. I suggest replacing "special relativity" by "the nonsignalling principle of special relativity".*
- *p3Lℓℓ30-31. It is not clear at this point that the (i, j, b) provided by each verifier to their prover can be different, i.e., that one verifier provides (i, j, b) to its prover whereas the other can provide a possibly different (i', j', b'). This becomes clear fairly quickly below, but at this point in reading the paper it is natural to incorrectly assume that the two challenges are identical because they are both called (i, j, b).*
- *p3Lℓ8. I would replace the word "ignore" by "take no account of" (just a suggestion).*
- *p3Lℓℓ37-39,43-44. The repetition of the fact that "the provers know that the graph is three-colourable" seems redundant to me.*
- *p3Rℓ34. Perhaps the title of this section should be "The graph" rather than "Graphs" since only one graph was used (again, just a suggestion).*
- *p3Rℓℓ34-42. This entire paragraph is very redundant with p1Rℓℓ1-10. It may be OK to repeat this material, but if you are short in space (for example to explain properly that the protocol given here is but a pedagogical variant of the one that was really implemented), this may be a good place to cut.*
- *p3Rℓ46. The words "algorithms efficient" should be swapped: "efficient algorithms".*
- *p3Rℓℓ55-56. You may wish to avoid this repetition of "so that".*
- *p4Rℓℓ10-11. Rather than saying that "A trigger signal is exchanged between the two verifiers", it would be more accurate to write that "A trigger signal is sent from the first verifier to the second".*
- *p5Lℓ5. Perhaps add "which is" in front of "completely". Also, perhaps say explicitly that this would require more than 3×10^15 rounds!*
- *p6Lℓ7. I see that there are two figures number 2: one above and one below the caption. Nevertheless, I find the expression "Figs 2" strange because of the plural Figs followed by the singular number 2.*
- *p6Lℓ23. Hmmm... Here we have that $a_1 + a_1' = [b_i r + c_i] + [b_i(-r) + c_i] \equiv 2c_i \pmod 3$ and similarly $a_2 + a_2' \equiv 2c_j \pmod 3$. It follows that it is correct to conclude that $c_i \neq c_j$ if (and only if) $a_1 + a_1' \not\equiv a_2 + a_2' \pmod 3$ as claimed, but only because $x = y \pmod 3 \Leftrightarrow 2x = 2y \pmod 3$. I am mentioning this because the main text may make the careless reader believe erroneously that $a_1 + a_1' = c_i$ and $a_2 + a_2' = c_j$.*

- *p6Rℓ7. Again, I would replace the word “ignore” by “take no account of”.*

* *p6Rℓ-20. I strongly object to writing “In Ref. [13] the authors give [. . .]” rather than naming said authors explicitly, especially in this case when there are only two authors.*

- *p8Lℓ27. I would replace the word “electronical” with “electronic”. When I checked for “electronical” with the help of Google, I was redirected to the Urban Dictionary, which says that this word is “a drunken reference to electronic equipment”. As for the word “electronicals” (which does not appear in the paper, but I could not resist the fun of mentioning it), it would be “anything using a power source not understood by the user”! ☺*

- *p8Lℓ34-35. This is slightly ambiguous. You mean that in each round both provers need to use the same random information, not that the same random information is used round after round!*

We thank the Referee for their minutious reading and helpful comments. We have implemented all these, which are marked in red in the resubmitted version.

Referee 3

We thank the Referee for their time and useful feedback. Below we provide point-by-point answers to each question or comment of the Referee.

This paper reports on an experimental demonstration of a well-known ZK proof of the second kind, the so-called GMW protocol from 1991 (Ref. [3]). The experiment involves setting up two distant communication links so that a round-trip communication can be performed at each distant space faster than the time it takes for light to travel between the two locations. As long as this can be ensured then the protocol is guaranteed to be sound: security rests on a “relativistic”, as opposed to “computational”, assumption.

From the theoretical point of view the theory used in the paper has been well-known since the early 1990s. From a practical point of view it is not clear if there is any real use for relativistic ZK protocols, because computational protocols are widely used and I don’t think there is any problem with them (e.g., the assumption of one-way functions is considered safe). Of course one could arguably say the same of quantum key distribution, so it is not a very fair criticism.

The issue with an assumption such as the existence of one-way functions is that, on the long run, it wears off. What is believed a solid candidate today may be broken twenty years from now. Moreover, a specific run of a protocol resting on a specific (believed) one-way function ultimately revolves around the difficulty of inverting a specific instance of a fixed size of the one-way function. This particular instance may turn out to be weaker than anticipated or just a very clever and long calculation may break it. The 1991 methodology à la GMW (Ref. [5], previously Ref. [3]) results in zero-knowledge proofs that completely lose their zero-knowledge aspect if the one-way function becomes broken. This means that an authentication protocol executed today may reveal the three-colouring of its graph five years from now because the one-way function involved will then be broken. Our protocol cannot be broken retroactively, no matter if information can later be transmitted faster than the speed of light. In our opinion, this represent a very significant advantage of relativistic zero-knowledge over existing methods.

For me the main interest of the paper lies in a "proof of principle" showing that protocols that may a priori seem completely out of reach can in fact be implemented using pretty straightforward equipment. An important drawback is that the protocol is not known to be secure in case the malicious parties may make use of quantum entanglement to cheat.

The case of quantum provers is indeed of great interest, and we have extended our discussion in the manuscript concerning this question.

It is indeed currently not known whether our protocol is secure against quantum provers. On this question, we share the point of view of Referee 2, namely that our current protocol could in fact be sound against quantum provers. This is supported by the fact that not a single graph that is not three-colourable (classically) has been found to be quantum three-colourable. More details on these questions, defining the notion of *quantum k-colourability* and existing examples can be found in Ref. [CMN+06].

However, and perhaps more importantly, there exist two adaptations of our protocol that are provably secure against quantum provers. First, by adding a third verifier-prover pair, following the method proposed in Ref. [24] (previously Ref. [22]). Second, by considering larger graphs with a number of vertices and edges quadratically bigger than the current number of vertices, following Ref. [25] (previously Ref. [23]). Unfortunately, at this stage, the implementation of either of these adaptations is out of reach in practice, as the required number of rounds is way too large. Nevertheless, there is clearly room for improvement on the security bounds, which may significantly reduce the number of rounds. Moreover, it may also be possible to combine both of these adaptations to obtain a more efficient protocol. As these extensions were not covered in details in the previous version of the manucsript, we have now written a separate section (Methods 5) to discuss these in detail.

In any case, if any change at all is required to guarantee security against quantum provers, it should definitely not be of paradigmatic nature. Both adaptations mentioned above do not modify the protocol per se, as they are forms of wrapping around our basic zero-knowledge protocol. Hopefully our work will motivate further theoretical investigations to find tighter bounds reducing the number of rounds within reasonable limits. A quadratic increase in the number of rounds instead of a fourth power would be quite manageable (see Methods 5).

In summary, the proof-of-principle is done: the ball is now in the hand of theorists.

I would like to know how the experimental setup differs from, or improves on, the one reported in [15]. It is known that ZK proofs can be constructed from secure bit commitments, so could the protocol from [15] not be adapted in a straightforward manner to obtain similar results as reported here?

This is an interesting comment. In fact, the starting point of our research was the observation that Ref. [17] (previously Ref. [15]) combined with Ref. [24] (previously Ref. [22]) does not yield a practical implementation of a zero-knowledge proof. The protocol of [24] uses the Hamiltonian cycle methodology of Blum. This means that for a graph of roughly 500 vertices, the provers will need 500*501/2 = 125 250 bit commitments executed in parallel à la [17] before the verifiers can disclose their decision to unveil the whole adjacency matrix or only a hamiltonian cycle. This requires a prohibitive amount of bandwidth between provers and verifiers. Moreover, the number of commitments grows quadratically with the number of vertices. If you use a graph with 1000 nodes, no communication equipment available today can handle the required bandwidth. The whole accomplishment of Ref. [14] (previously Ref. [12]) was to reduce the total number of commitments necessary down to four per iteration. Sustaining commitments in the manner of [17] is extremely expensive in terms of bandwidth. In contrast, our protocol can handle graphs of 1000 nodes with nearly no extra effort. That's because the total communication cost of our protocol grows logarithmically with the number of vertices and edges. A straightforward combination of Ref. [17] with Ref. [5] (GMW, previously Ref. [3]) still grows linearly with the number of edges.

Additional comments:

- *In the first page, it would be worth explaining the notion of a one way function when first introducing it, as this may not be part of the basic knowledge of a Nature reader.*

In the main text, the first occurrence of the terminology is indeed accompanied by a small explanation: "one-way functions, that is, functions that can be efficiently computed but for which finding a preimage of a particular output cannot".

As for the abstract, we considered implementing the Referee's remark but the resulting sentence was too verbose to meet the conciseness needed there. Thus we would prefer to keep the formulation as it was.

- *p.3 1st column: "pre-agree on random three-colourings", the colourings should be proper.*

We agree with the Referee that our use of the word "colouring" was not completely rigourous. We have now added a sentence in the section *Protocol* (page 3) to mention that we always imply this property when using this word. As for *improper* colourings, we employ the term "labelling" instead of "colouring" to avoid confusion for a reader not accustomed to this subtle distinction.

- *Your protocol makes use of two verifiers and two provers. This is somewhat unexpected, as usually ZK proofs involve a single verifier and two provers. Could you explain why? Does the use of two verifiers make the experiment more accessible?*

Special relativity is developed around the principle that communication cannot happen faster than the speed of light. In the setting of relativistic zero-knowledge proofs, not only are there numerous provers but also several verifiers. In order to guarantee some distance between the provers we can position verifiers at known distance from each other and measure the response time of the provers. By doing so they can validate that the provers cannot possibly have communicated among themselves because they responded fast enough *wherever they were* located. Therefore the verifiers do not need to precisely locate the provers. In contrast, the theoretical Multi-prover Interactive Proof model with a single verifier does not state *how* it will enforce no-signalling between the provers. As Referee 2 also raised a related question, we added an explanatory sentence on page 2.

References

[CMN+06] Peter J. Cameron, Ashley Montanaro, Michael W. Newman, Simone Severini, and Andreas Winter. On the quantum chromatic number of a graph. *arXiv:quant-ph/0608016*, 2006.

[CY19] Claude Crépeau and Nan Yang. Non-Locality and Zero-Knowledge **MIP**s. *arXiv:1907.12619*, 2019.

Reviewer Reports on the First Revision:

Referee #3:
Remarks to the Author:
The updated manuscript is well-written, and more accessible than the earlier version; all points recommended by the referees have been taken into account.

Since my previous opinion was to publish the paper in a more specialized journal, and the results are unchanged, I maintain that opinion.

Referee #2:
Remarks to the Author:

Referee Report on **revised** version of *Experimental relativistic zero-knowledge proofs*

Nature manuscript 2021-01-00724

Referee report dated 6 June 2021

The authors of this paper have taken very seriously my original referee report and I find their answers to the two other referee reports perfectly adequate. Furthermore, I commend their extended discussion on quantum provers (Methods 5). Therefore, I am more than ever enthusiastic in my recommendation that you accept their paper for a well-deserved publication in *Nature*. Nevertheless, two final corrections are required before publication and I have a few additional suggestions that the authors may wish to consider (or not).

My main recommendation concerning the original paper was to avoid describing two different protocols, a simple one explained in the main text and a more complicated one (the original protocol from Ref [14], which was the one they had actually implemented, I assume before discovering the simplified protocol) in the Methods. They followed this recommendation beyond my expectations by performing their experiment all over again with the simplified protocol, which allowed then to describe only that protocol in both the main text (as a very brief sketch) and the Methods (in details). Unfortunately, they left vestiges of the original protocol from Ref [14] in their detailed description of the simplified protocol in the Methods! Specifically, in the last paragraph of the left column of page 6, they kept the *randomizers* (r, s) from the original protocol rather than the single bit b that occurs in the simplified protocol. Also, the two possible actions that the verifiers should take with probability $2/5$ differ in whether it is the first or the second randomizer that should be the same for both verifiers, which of course makes no sense whatsoever now that there is a single bit b in the protocol rather than two randomizers r and s. I am sure this paragraph can be fixed since in fact this had to be the case in the actual revised implementation. This issue *must* be corrected in the Methods before the paper can be published.

Furthermore, the authors made a modification that, as far as I can see, was not suggested by any of the three referees (certainly not by me), and in my opinion that was erroneous. In the originally submitted manuscript, the last sentence in the caption of Figure 1 was "even with all the provers' answers at hand, the verifiers are not more efficient at elaborating a three-colouring than initially (zero-knowledge)" whereas in the revised version it is "even with all the provers' answers in hand, the verifiers are not more efficient at deciding three-colourability than initially (zero-knowledge)". Well, the new statement is false. Now, the verifiers *are* much more efficient at *deciding* if the graph of interest is three-colourable: they have learned beyond any reasonable doubt that it is (unless they caught the provers cheating)! What they *are* still not able to do more efficiently than before is *find* (aka "elaborate") a three-colouring of the graph of interest, which is precisely what the old sentence said. I am completely puzzled by what made them commit this change in their caption. (As a side remark, they were right in changing "at hand" into "in hand", but the fact that they did not set the word "in" in red indicates that they cannot be trusted in their claim that their changes are "all marked in red in the manuscript".)

Other than these two major points, I noticed a few minor ones on which I do not insist in the least. Please note that I did not re-read the entire paper but rather concentrated on changes that I requested in my original report (all is fine on this front unless otherwise noted) and on the portions of the revised text they set in red, under the act of faith that this is indeed all that was changed between the two versions even though I know from the parenthesis at the end of the previous paragraph that this is not strictly the case.

As I did in my report on the original submission, I refer to the page number ("p"), the side ("L" for left of "R" for right) and the line number ("ℓ") according to the revised manuscript. The absolute value of negative line numbers should be counted up from the bottom of the page.

Specific Minor Comments

- $p2$, caption of FIG. 1 and several other places. In my opinion, it would look nicer to use `\ell` in constructions such as ℓ_k^0, ℓ_k^1 and (later) ℓ_i^b, etc.
- $p3\mathsf{L}\ell28$. I think there is something grammatically incorrect in the expression "to [...] answer to the verifiers".
- $p3\mathsf{L}\ell32$. The amount of spacing after the (red) full stop is excessive.
- $p4\mathsf{L}\ell6$. The phrase "to the prover it is connected to" can be criticised (grammatically speaking). I think "to the prover to which it is connected" would be preferable.
- $p4\mathsf{L}\ell16$. I am curious as to why the repetition rate went from 3 MHz to 0.5 MHz between the experiment described in the original paper and the revised one. Was the original hardware no longer available or is there a more fundamental reason? Obviously, there is no need for the authors to address this issue in the final version.
- $p4\mathsf{L}\ell17$. Given that your repetition rate has slowed down by a factor of 6, how is it possible that what used to take "less than a second" in the original paper now takes "two seconds" (rather than "less than six seconds")? Well, of course if it took $^2/_6 = {}^1/_3$ second in the original experiment, it *was* "less than a second", but you could have boasted better in that case. Furthermore, the fact that $^1/_3 < 1$ cannot explain the next point.

- $p4\mathsf{L}\ell 22$. It was “about 3s” with the repetition rate of 3 MHz. Why is it not “about 18 seconds” (rather than “about ten seconds” with the repetition rate of 0.5 MHz?

- $p4\mathsf{R}\ell 8$. I did not “complain” about this in the original submission even though I winced, but in my opinion there is no such thing as “quantum nonlocality”. Why not replace “a phenomenon known as quantum nonlocality” by “a phenomenon due to quantum entanglement”? Of course, Ref. [22] may become irrelevant.

- $p5\mathsf{L}\ell{-}11$. Rather than writing “l_i^b and l_j^b”, I suggest writing (l_i^b, l_j^b) or, better, (ℓ_i^b, ℓ_j^b), and the same remark applies to what should be $(\ell_{i'}^{b'}, \ell_{j'}^{b'})$ on the next line. The reason is that on the last line of this column you mention “the answers (a_1, a_2) and (a_1', a_2')”. Does each prover return two one-trit answers or a single answer, which happens to be a pair of qutrits? My suggestion would make it easier to connect the dots.

- $p6$, caption of FIG. 2. After “to the provers”, I suggest adding “but $b \neq b'$”.

- $p6\mathsf{L}\ell{-}14$. The repetition of the word “following’” is awkward

- $p6\mathsf{R}\ell 12$. In my opinion, this “which” should be “that”.

- $p8\mathsf{R}\ell 12$. You need a comma before this “which”.

- $p8\mathsf{R}\ell{-}16$. I think “commutating” should be “commuting”.

- $p8\mathsf{R}\ell{-}3$. Another way to “bring the number of rounds down to something practical” would be to prove that the protocol you implemented does not need to be replaced by something more complicated because it is quantum-safe already.

Referee 2

The authors of this paper have taken very seriously my original referee report and I find their answers to the two other referee reports perfectly adequate. Furthermore, I commend their extended discussion on quantum provers (Methods 5). Therefore, I am more than ever enthusiastic in my recommendation that you accept their paper for a well-deserved publication in Nature.

We thank the Referee for their enthusiasm and further useful comments.

Nevertheless, two final corrections are required before publication and I have a few additional suggestions that the authors may wish to consider (or not).

My main recommendation concerning the original paper was to avoid describing two different protocols, a simple one explained in the main text and a more complicated one (the original protocol from Ref [14], which was the one they had actually implemented, I assume before discovering the simplified protocol) in the Methods. They followed this recommendation beyond my expectations by performing their experiment all over again with the simplified protocol, which allowed then to describe only that protocol in both the main text (as a very brief sketch) and the Methods (in details). Unfortunately, they left vestiges of the original protocol from Ref [14] in their detailed description of the simplified protocol in the Methods! Specifically, in the last paragraph of the left column of page 6, they kept the randomizers (r, s) from the original protocol rather than the single bit b that occurs in the simplified protocol. Also, the two possible actions that the verifiers should take with probability $\frac{2}{5}$ differ in whether it is the first or the second randomizer that should be the same for both verifiers, which of course makes no sense whatsoever now that there is a single bit b in the protocol rather than two randomizers r and s. I am sure this paragraph can be fixed since in fact this had to be the case in the actual revised implementation. This issue must be corrected in the Methods before the paper can be published.

We thank the Referee for their careful reading and we apologise for these relics that we failed to edit consistently. The new version naturally takes this comment into account.

Furthermore, the authors made a modification that, as far as I can see, was not suggested by any of the three referees (certainly not by me), and in my opinion that was erroneous. In the originally submitted manuscript, the last sentence in the caption of Figure 1 was "even with all the provers answers at hand, the verifiers are not more efficient at elaborating a three-colouring than initially (zero-knowledge)" whereas in the revised version it is "even with all the provers answers in hand, the verifiers are not more efficient at deciding three-colourability than initially (zero-knowledge)". Well, the new statement is false. Now, the verifiers are much more efficient at deciding if the graph of interest is three-colourable: they have learned beyond any reasonable doubt that it is (unless they caught the provers cheating)! What they are still not able to do more efficiently than before is find (aka "elaborate") a three-colouring of the graph of interest, which is precisely what the old sentence said. I am completely puzzled by what made them commit this change in their caption.

This remark is absolutely correct and we have reversed the scientific content of the sentence to its original version.

- *p2, caption of FIG. 1 and several other places. In my opinion, it would look nicer to use* `\ell` *in constructions such as ℓ_k^0, ℓ_k^1 and (later) ℓ_i^b, etc.*

We do agree with this comment and have implemented it consistently through the article.

- *p4Lℓ16. I am curious as to why the repetition rate went from 3 MHz to 0.5 MHz between the experiment described in the original paper and the revised one. Was the original hardware no longer available or is there a more fundamental reason? Obviously, there is no need for the authors to address this issue in the final version.*

The hardware used in both cases was almost exactly the same. The reason for the change is different and originates from a misunderstanding between the authors while writing the first version of the manuscript. Having ran the experiment again allowed us to notice this error and correct it.

- *p4Lℓ17. Given that your repetition rate has slowed down by a factor of 6, how is it possible that what used to take "less than a second" in the original paper now takes "two seconds" (rather than "less than six seconds")? Well, of course if it took $\frac{2}{6} = \frac{1}{3}$ second in the original experiment, it was "less than a second", but you could have boasted better in that case. Furthermore, the fact that $\frac{1}{3} < 1$ cannot explain the next point.*

In the first version, we indeed thought to have $\frac{1}{3}$ second in the triggered version. Given that the final message of the article is that more effort could be devoted to reduce the distance and the time needed, we decided to only give a rough idea of the time involved in our protocol. Therefore, "less than a second" was convincing enough to our eyes.

- *p4L‘22. It was “about 3s” with the repetition rate of 3 MHz. Why is it not “about 18 seconds” (rather than “about ten seconds”) with the repetition rate of 0.5 MHz?*

This number was indeed incoherent in its context. This originates from a minor theoretical improvement that happened roughly at the same time as the resubmission. Specifically, the number of rounds decreased from $9|E|k$ to $5|E|k$, which now appears in the text, so that the current number is correct. We are sorry for the inconsistency of the previous version; extra care has been taken for the new one.

- *p4Rℓ8. I did not “complain” about this in the original submission even though I winced, but in my opinion there is no such thing as “quantum nonlocality”. Why not replace “a phenomenon known as quantum nonlocality” by “a phenomenon due to quantum entanglement”? Of course, Ref. [22] may become irrelevant.*

We have taken this remark into account.

- *p5Lℓ-11. Rather than writing “ l_i^b and l_j^b ”, I suggest writing (l_i^b, l_j^b) or, better, (ℓ_i^b, ℓ_j^b), and the same remark applies to what should be $(\ell_{i'}^{b'}, \ell_{j'}^{b'})$ on the next line. The reason is that on the last line of this column you mention “the answers (a_1, a_2) and (a_1', a_2')”. Does each prover return two one-trit answers or a single answer, which happens to be a pair of qutrits? My suggestion would make it easier to connect the dots.*

We agree with the comment and have edited the manuscript accordingly. Each prover gives a single answer, which happens to be a pair of qutrits.

- *p3Lℓ28. I think there is something grammatically incorrect in the expression “to [. . .] answer to the verifiers”.*
- *p3Lℓ32. The amount of spacing after the (red) full stop is excessive.*
- *p4Lℓ6. The phrase “to the prover it is connected to” can be criticised (grammatically speaking). I think “to the prover to which it is connected” would be preferable.*
- *p6, caption of FIG. 2. After “to the provers”, I suggest adding “but $b \neq b'$ ”.*
- *p6Lℓ-14. The repetition of the word “following” is awkward*
- *p6Rℓ12. In my opinion, this “which” should be “that”.*
- *p8Rℓ12. You need a comma before this “which”.*
- *p8Rℓ-16. I think “commutating” should be “commuting”.*

We have taken into account all of these minor points and we thank again the Referee for their minuteness.

- *p8Rℓ-3. Another way to "bring the number of rounds down to something practical" would be to prove that the protocol you implemented does not need to be replaced by something more complicated because it is quantum-safe already.*

We have added a sentence in this way at the very end of the article.