A proposal to get some common-sense intuition for the paradox of the double-slit experiment

Gerrit Coddens, Andre Gontijo Campos

▶ To cite this version:

Gerrit Coddens, Andre Gontijo Campos. A proposal to get some common-sense intuition for the paradox of the double-slit experiment. 2016. cea-01383609v2

HAL Id: cea-01383609

https://cea.hal.science/cea-01383609v2

Preprint submitted on 26 Oct 2016 (v2), last revised 26 Oct 2021 (v10)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

epjp manuscript No.
(will be inserted by the editor)

A proposal to get some common-sense intuition for the paradox of the double-slit experiment

Gerrit Coddens (1) and Andre Gontijo Campos (2)

(1) Laboratoire des Solides Irradiés, Ecole Polytechnique, CNRS, CEA, Université Paris-Saclay, 91128-Palaiseau CEDEX, FRANCE
(2) Department of Chemistry, Princeton University, Princeton, NJ 08544, USA

Abstract. We argue that the double-slit experiment can be understood much better by considering it as an experiment whereby one uses electrons to study the set-up rather than an experiment whereby we use a set-up to study the behaviour of electrons. We also show that Heisenberg's uncertainty principle is related to Gödel's concept of undecidability and how the latter can be used in an intuitive way to make sense of the double-slit experiment and the quantum rules for calculating coherent and incoherent probabilities. We meet here a situation where the electrons always behave in a fully deterministic way, while the detailed design of the set-up may render the question about the way they move through the set-up experimentally undecidable. Heisenberg's uncertainty relation is a rule of thumb to predict such undecidability (but only one among others). It is very important to make a distinction in quantum mechanics between the determinism of nature (Einstein) and the decidability of a question within an experimental set-up (Bohr). The former is about the absolute truth of an answer to a yes or no question, and follows binary logic (true or false), the latter about what an experimental set-up can decide and tell about the truth of that answer, and follows ternary logic (true, false or undecidable). Binary and ternary logic are incompatible. The viewpoints of Bohr and Einstein are thus operating on different levels and it is only by confusing these two levels that these two viewpoints seem to be irreconcilable. A very important element in the analysis is the problem of the existence of *common* probability distributions. And this is a recurrent theme in other situations that are felt as paradoxical. The CHSH inequality used in the experiments of Aspect *et al.* clashes with quantum mechanics because it is based on the assumption that there would exist a *common* probability distribution for the hidden variables in the different experiments that intervene in the inequality. That such a common probability distribution does not exist is well known from quantum mechanics (because the operators that come into play do not commute and therefore do not have common eigenstates, i.e. common probability amplitudes). But this fact is not a prerogative of quantum mechanics and can also be explained by purely classical reasoning.

PACS. 03.65.Ta Foundations of Quantum Mechanics; measurement theory – 42.25.Hz Interference – 02.20.-a Group Theory – 03.65.Pm Relativistic wave equations

1 Introduction

In reference [1] we derived the Dirac equation by just expressing that the electron is a spinning particle.[1] The derivation is entirely classical in the sense that we start by considering a particle in free-space that travels in uniform motion on a straight line according to $\mathbf{r}(t) = \mathbf{r}_0 + \mathbf{v}t$. A Dirac-like equation (which is still different from the Dirac equation) is then derived on the set $\mathcal{P} = \{(\mathbf{r}, t) \in \mathbb{R}^4 \ \| \ \mathbf{r}(t) = \mathbf{r}_0 + \mathbf{v}t\}$. A last step in the derivation (which we will explain below in Section 2) permits to obtain on $\mathcal{P}$ the Dirac equation from the Dirac-like equation. On reading these lines one may think that it cannot possibly be right that the quantum mechanical Dirac equation can be derived from a classical reasoning: It just sounds crank. However, we have argued in reference [1] (see p. 125) and we will illustrate in this paper how *it is not the way we derive it but the way we solve it, that turns the Dirac equation into quantum mechanics*. In fact, when we solve the wave equations in quantum mechanics we transgress the definition domain of the classically derived equations in two instances.[2]

The first one occurs when we derive the Dirac equation from the Dirac-like equation. In this step, we replace a spinor by a sum of spinors. While it is algebraically perfectly feasible to make a linear combination of two spinors, the meaning of such a procedure is not defined by the group theory. The reason for this is that spinors represent group elements and that the rotation

[1] The current dogma is that spin does not correspond to spinning motion because such an assumption cannot explain the magnetic moment of the electron. But this conclusion is not compelling as discussed in [2].

[2] Throughout the paper, we rely on the fact that the Schrödinger can be derived from the Dirac equation, such that we cover both equations by discussing everything in terms of the Dirac equation.

group and the homogeneous Lorentz group we use in physics are not vector spaces but curved Lie group manifolds.[3] Making a linear combination of spinors is thus *a priori* a meaningless operation within the pristine conceptual frame work of the group theory.[4] One could qualify using such linear combinations as a mindless use of the algebra whereby one does not realize that one transgresses the limits of the special definition domain and conceptual framework of the group theory in carrying out such calculations. In doing so, one complies exactly with the motto: "shut up and calculate". Fortunately, one can find an *a posteriori* justification for this procedure, by finding a meaning for it in terms of sets. These sets can be seen to represent statistical ensembles and their use turns therefore the Dirac equation into statistical physics. We will explain this in Section 2.

The second transgression occurs when we solve the Dirac equation. In fact, following the classical ideas that underly the derivation, we should solve the Dirac equation on the set $\mathcal{P}$. Instead of that, we solve it generously and simplemindedly over $\mathbb{R}^4$, without realizing that we only needed to solve it over $\mathcal{P}$. While solving the equation over $\mathbb{R}^4$ is algebraically possible, it is once more a shut-up-and-calculate approach of which the meaning has *a priori* not been defined, such that we must again try to find an *a posteriori* justification for it. As we claimed that it is the way we solve the Dirac equation that turns it into quantum mechanics, we can anticipate already that this will prove a very difficult task. And as a matter of fact, it is through this kind of over-zealous extrapolation of the definition domain of the equation from $\mathcal{P}$ to $\mathbb{R}^4$ that we allow all the trickery of quantum mechanics to enter the scene.[5] It is thus very important to figure out what this extrapolation exactly means. We will see that there is no unique answer to this question and that the solution may very much depend on the specific physical context considered, such that it requires a discussion on a case-by-case basis.

It is the aim of this paper to try to clarify the issues which arise in these two transgressions. We will discuss them at the hand of a result of quantum mechanics that is particularly counter-intuitive and difficult to understand, viz. the double-slit experiment. The discussion about the two transgressions of the definition domains in the mathematics is of a type that may look like splitting hairs to many a physicist. But they really lead straight into the heart of the matter. The analysis we present here of the double-slit experiment is meant to show, we hope in a convincing way, the truth of the daring thesis that the Dirac equation is classical and that it is only the way we solve it that renders it quantum mechanical. *The readers who want to make the effort to study both the derivation of the Dirac equation in reference [1] and the present paper, will be able to verify that the whole presents a rigorous (but rather long) step-by-step mathematical argument that starts from scratch and whereby every step is justified and intuitive.*

We hope the reader will recognize the importance of this approach. Its power resides in the fact that one can exactly spot where and how we cross the frontier between classical mechanics and quantum mechanics. We can use the knowledge about the exact location of this frontier as a bistoury to analyze in detail what the transgressions may mean and this way gain a better understanding of quantum mechanics. This is something that we may never have guessed or anticipated, because at face value it

[3] E.g. in SU(2), spinors represent rotations as explained in references [1] (see pp. 48-52) and [2]. We can then get a feeling for the fact that a linear combination of two spinors will not be a new spinor by an analogy. A linear combination of two rotation matrices in SO(3) is not a new rotation matrix.

[4] The example of 3 × 3 rotations matrices suggests that we could use it as a source of inspiration to give a meaning to a linear combination $\mathbf{M} = c_1\mathbf{R}_1 + c_2\mathbf{R}_2$ of rotation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. It is the matrix that transforms a vector $\mathbf{v}$ expressed as the 3×1 column matrix $[\,\mathbf{v}\,]$ to $c_1[\,\mathbf{v}_1\,] + c_2[\,\mathbf{v}_2\,]$ whereby $[\,\mathbf{v}_1\,] = \mathbf{R}_1[\,\mathbf{v}\,]$ and $[\,\mathbf{v}_2\,] = \mathbf{R}_2[\,\mathbf{v}\,]$. This is a valid procedure for the 3×3 representation, because the 3×1 matrices represent elements of a vector space. But in SU(2), the spinors are just a stenographic notation for the SU(2) matrices obtained by taking their first columns (see reference [1]). The procedure to give meaning to linear combinations of SU(2) matrices in analogy with the procedure in SO(3) can then not be used because the second-stage question what a linear combination of spinors would be is exactly the same question as the original question what a linear combination of SU(2) matrices would be. This turns the attempt to solve the problem in analogy with SO(3) into a vicious circle. The analogous procedure fails because spinors (which are the 2×1 matrices) do not belong to a vector space like the vectors of $\mathbb{R}^3$ (which are the 3×1 matrices). We could of course argue that we must search for the meaning of the sum of spinors by translating the problem to $\mathbb{R}^3$, solve it there, and then translate the solution backwards to $\mathbb{C}^2$, but this has several problems in its own right:

(1) First of all it leads exactly to the conceptual difficulty that in Atiyah's words: "A spinor is the square root of a vector" [3]. An illustration of this fact is that probabilities (which from the relativistic point of view are the time-component of a probability charge-current four-vector) are squares of probability amplitudes ψ (which relativistically are also a component of a four-component spinor-like quantity Ψ). Understanding the difference between summing probabilities and summing probability amplitudes is the very hard nut we have to crack if we want to make sense of the double-slit experiment, which in Feynman's words is the only mystery of quantum mechanics [4]. As this quadratic relationship is very difficult to understand, the clear meaning of the sum of vectors in SO(3) gets lost in the attempt to translate it backwards to SU(2) and to solve the riddle of the meaning of summing probability amplitudes. The relation between the SO(3) matrices and the SU(2) matrices is also quadratic.

(2) The second problem is that we are straying out of the conceptual framework we set up to study the group. This conceptual framework of SU(2) contains all we need to know about the group in a self-contained way. We should thus not look for solutions elsewhere. It is completely artificial to draw in *external* considerations from SO(3) into SU(2). By external considerations we mean here that we are using arguments from another representation SO(3) to settle problems in SU(2). The logic of SU(2) should be self-contained (i.e. internal) and not rely on arguments drawn in from SO(3). Moreover, the *internal* approach we will develop and which is based on sets and statistical ensembles reproduces exactly (within a completely self-consistent internal logic) the quantum rules that have already been proposed in physics to solve the riddle how we must interpret a superposition of two states. These rules can be formulated within SU(2) without invoking SO(3) at any time.

[5] This can already be appreciated from the fact that we solve the Dirac or the Schrödinger equation also over the classically forbidden regions under and at the other side of the barrier when we solve these equations for a tunneling experiment. It is certainly legitimate to ask why we should consider such a startling extrapolation of the definition domain.

just does not sound right to claim that the Dirac equation can be derived classically. We must also point out that in the traditional approach to quantum mechanics, one cannot possibly become aware of the two transgressions we are discussing here. They become only apparent in our approach, which may illustrate why it could be important. Any possible perceived irony about these transgressions is thus just for didactical reasons. It may already have transpired from the preceding lines, but we must warn the reader that it requires a special mindset to meet with our approach. Its merits cannot be evaluated by promoting (a plethora of) beliefs anchored in the traditional approach to peremptory touchstones. That would be more or less like criticizing hyperbolic geometry from the stronghold of the premisses of Euclidean geometry.

2 The first transgression: From a Dirac-like to the Dirac equation

With the Cartan-Weyl choice for the gamma matrices, the free-space Dirac-like equation reads:

$$\begin{bmatrix} & -\frac{\hbar}{\imath}\frac{\partial}{\partial ct}\,\mathbb{1} - \frac{\hbar}{\imath}\boldsymbol{\nabla}_{\|}{\cdot}\boldsymbol{\sigma} \\ -\frac{\hbar}{\imath}\frac{\partial}{\partial ct}\,\mathbb{1} + \frac{\hbar}{\imath}\boldsymbol{\nabla}_{\|}{\cdot}\boldsymbol{\sigma} & \end{bmatrix}\begin{bmatrix} \varPsi & \\ & \varPsi^{-1\dagger} \end{bmatrix} = m_0 c\begin{bmatrix} & s_{ct}\mathbb{1} + \mathbf{s}{\cdot}\boldsymbol{\sigma} \\ -(s_{ct}\mathbb{1} - \mathbf{s}{\cdot}\boldsymbol{\sigma}) & \end{bmatrix}\begin{bmatrix} \varPsi & \\ & \varPsi^{-1\dagger} \end{bmatrix}. \tag{1}$$

Here $\mathbf{s}$ is the spin axis. Very often one supposes that $\mathbf{s} = \mathbf{e}_z$ as we identify $\mathbf{s}$ with the basis vector $\mathbf{e}'_z$ of some triad of basis vectors that have been originally attached to the electron to permit following its rotational motion. The directional partial derivative $\boldsymbol{\nabla}_{\|}$ is here taken in the direction of a boost velocity vector $\mathbf{v}$. The fact that $-(s_{ct}\mathbb{1} - \mathbf{s}{\cdot}\boldsymbol{\sigma})$ occurs here with a minus sign is due to the fact that $\mathbf{s}$ is a pseudo-vector. This equation is derived from the Rodrigues formula in SU(2) for a rotation around an axis $\mathbf{n}$ over an angle φ within a frame at rest:

$$\mathbf{R}(\mathbf{n},\varphi) = \cos\frac{\varphi}{2} - \imath\sin\frac{\varphi}{2}\,[\,\mathbf{n}{\cdot}\boldsymbol{\sigma}\,]. \tag{2}$$

The substitution $\varphi\,|\,\omega_0\tau$ (where τ is the proper time) transforms this into the description of a particle that spins around $\mathbf{n}$. After using the 2×1 spinor ψ as a stenographic notation for the 2×2 matrix $\mathbf{R}(\mathbf{n},\varphi)$ by taking its first column, and taking the time derivative, one obtains:

$$\frac{d}{d\tau}\psi(\tau) = -\imath\frac{\omega_0}{2}\,[\,\mathbf{n}{\cdot}\boldsymbol{\sigma}\,]\,\psi(0). \tag{3}$$

The fact that we use $\mathbf{s}$ instead of $\mathbf{n}$ in Eq. 1 hides a tedious technicality we discuss in reference [1]. The idea is to render the equation independent from the choice of a reference frame by rendering it covariant. This does not work with $\mathbf{n}$ because it is not a covariant quantity and it does not transform like a vector, while it works for $\mathbf{s}$ which is a covariant quantity and does transform like a vector. Eq. 1 is just a covariant reformulation of Eq. 3 in a moving frame with boost velocity $\mathbf{v}$. We have then $(\frac{d}{dc\tau}, 0) \to (\frac{\partial}{\partial ct}, \nabla_{\|})$. To obtain Eq. 1 from Eq. 3, one must introduce $m_0c^2 = \hbar\omega_0/2$ (with the effect that $e^{-\imath\omega_0\tau/2}$ becomes $e^{-\frac{\imath}{\hbar}(Et-\mathbf{p}\cdot\mathbf{r})}$), and lift the equation from SU(2) to the representation of the Lorentz group in terms of gamma matrices. All these steps are discussed in full detail in reference [1]. We obtain this way the description of a spinning particle that moves at a constant velocity $\mathbf{v}$ along a straight line $\mathbf{r} = \mathbf{r}_0 + \lambda\mathbf{v}$, $\lambda \in \mathbb{R}$. If we can assume that $\mathbf{s} \perp \mathbf{v}$ then $s_t = 0$ and:

$$\begin{bmatrix} & -\frac{\hbar}{\imath}\frac{\partial}{\partial ct}\,\mathbb{1} - \frac{\hbar}{\imath}\boldsymbol{\nabla}_{\|}{\cdot}\boldsymbol{\sigma} \\ -\frac{\hbar}{\imath}\frac{\partial}{\partial ct}\,\mathbb{1} + \frac{\hbar}{\imath}\boldsymbol{\nabla}_{\|}{\cdot}\boldsymbol{\sigma} & \end{bmatrix}\begin{bmatrix} \varPsi & \\ & \varPsi^{-1\dagger} \end{bmatrix} = m_0 c\begin{bmatrix} & \mathbf{s}{\cdot}\boldsymbol{\sigma} \\ \mathbf{s}{\cdot}\boldsymbol{\sigma} & \end{bmatrix}\begin{bmatrix} \varPsi & \\ & \varPsi^{-1\dagger} \end{bmatrix}. \tag{4}$$

The fact that $\mathbf{s}$ is an axial vector is even more obvious in this equation. We have now arrived at the point where we must transgress the definition domain of the algebra. To obtain the Dirac equation from Eq. 4, one would need that:

$$\begin{bmatrix} & \mathbf{s}{\cdot}\boldsymbol{\sigma} \\ \mathbf{s}{\cdot}\boldsymbol{\sigma} & \end{bmatrix}\begin{bmatrix} \varPsi & \\ & \varPsi^{-1\dagger} \end{bmatrix} = \begin{bmatrix} \varPsi & \\ & \varPsi^{-1\dagger} \end{bmatrix}. \tag{5}$$

But this is just not possible. An operator:

$$\mathbf{S} = \begin{bmatrix} & \mathbf{s}{\cdot}\boldsymbol{\sigma} \\ \mathbf{s}{\cdot}\boldsymbol{\sigma} & \end{bmatrix} \tag{6}$$

can after operating on a group element:

$$\boldsymbol{\varPsi}_1 = \begin{bmatrix} \varPsi & \\ & \varPsi^{-1\dagger} \end{bmatrix} \tag{7}$$

never yield the group element $\boldsymbol{\varPsi}_1$ again because $\mathbf{S}\boldsymbol{\varPsi}_1$ must have an off-diagonal block structure while $\boldsymbol{\varPsi}_1$ has a diagonal block structure. (Matrices representing group elements obtained from an even number of reflections, i.e. rotations and boosts, always have a block-diagonal structure in the Cartan-Weyl choice for the gamma matrices). To obtain the Dirac equation, we must thus introduce a second quantity:

$$\boldsymbol{\Psi}_2 = \left[\begin{array}{cc} & \mathbf{s}{\cdot}\boldsymbol{\sigma} \\ \mathbf{s}{\cdot}\boldsymbol{\sigma} & \end{array}\right]\left[\begin{array}{cc} \Psi & \\ & \Psi^{-1\dagger} \end{array}\right] = \mathbf{S}\boldsymbol{\Psi}_1, \tag{8}$$

and introduce a superposition of states: $\boldsymbol{\Phi} = \boldsymbol{\Psi}_1 + \boldsymbol{\Psi}_2$:

$$\left[\begin{array}{cc} \Psi & \\ & \Psi^{-1\dagger} \end{array}\right] + \left[\begin{array}{cc} & \mathbf{s}{\cdot}\boldsymbol{\sigma} \\ \mathbf{s}{\cdot}\boldsymbol{\sigma} & \end{array}\right]\left[\begin{array}{cc} \Psi & \\ & \Psi^{-1\dagger} \end{array}\right] = \left[\begin{array}{cc} \Psi & [\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]\,\Psi^{-1\dagger} \\ [\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]\,\Psi & \Psi^{-1\dagger} \end{array}\right] = \boldsymbol{\Phi}, \tag{9}$$

such that $\mathbf{S}\boldsymbol{\Phi} = \mathbf{S}(\boldsymbol{\Psi}_1 + \boldsymbol{\Psi}_2) = \mathbf{S}(\boldsymbol{\Psi}_1 + \mathbf{S}\boldsymbol{\Psi}_1) = \mathbf{S}\boldsymbol{\Psi}_1 + \mathbf{S}^2\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2 + \boldsymbol{\Psi}_1 = \boldsymbol{\Phi}$ and:

$$\left[\begin{array}{cc} & -\frac{\hbar}{\imath}\frac{\partial}{\partial ct}\,\mathbb{1} - \frac{\hbar}{\imath}\boldsymbol{\nabla}_{\|}{\cdot}\boldsymbol{\sigma} \\ -\frac{\hbar}{\imath}\frac{\partial}{\partial ct}\,\mathbb{1} + \frac{\hbar}{\imath}\boldsymbol{\nabla}_{\|}{\cdot}\boldsymbol{\sigma} & \end{array}\right]\boldsymbol{\Phi} = m_0 c\,\boldsymbol{\Phi}. \tag{10}$$

This is algebraically feasible, but geometrically meaningless. First of all, it is even not clear that $\mathbf{S}$ is a group element, but even if it is, such that $\boldsymbol{\Psi}_2$ represents a group element, a quantity $\boldsymbol{\Psi}_1 + \boldsymbol{\Psi}_2$ is *a priori* not defined as we already mentioned in the Introduction.

In fact, in group representation theory the meaning of $\mathbf{D}(g_2)\,\mathbf{D}(g_1) = \mathbf{D}(g_2 \circ g_1)$, where the representation matrices $\mathbf{D}(g_j)$ represent the group elements $g_j \in G$ and $\circ$ represents the group operation of the group $(G, \circ)$, is well defined. But the meaning of a linear combination $c_2\mathbf{D}(g_2) + c_1\mathbf{D}(g_1)$, where $(c_1, c_2) \in \mathbb{C}^2$ or $(c_1, c_2) \in \mathbb{R}^2$ is not defined. The tentative homomorphism between $c_2\mathbf{D}(g_2) + c_1\mathbf{D}(g_1)$ and $c_1 g_1 + c_2 g_2$ has not been defined, and even the operation $c_1 g_1 + c_2 g_2$ itself on the group has *a priori* not been defined. This leads us straight into the thorny problems evoked in Footnote 4.

We must thus try to give a meaning to the undefined procedure of calculating $\boldsymbol{\Phi} = \boldsymbol{\Psi}_1 + \boldsymbol{\Psi}_2$ and more generally $c_2\mathbf{D}(g_2) + c_1\mathbf{D}(g_1)$. We will define this way, what has been called the group ring [5]. We can illustrate the idea for Pauli's formalism for the spin in SU(2). For a reflection operator $[\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]$ we can consider the eigenvalue equation:

$$[\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]\,\psi = \lambda\psi. \tag{11}$$

Here $\mathbf{s}$ is a unit vector, while ψ is again a 2×1 spinor. The unit vector $\mathbf{s}$ defines the reflection operator, whose reflection plane is orthogonal to $\mathbf{s}$. This reflection operator is proportional to the spin operator $\frac{\hbar}{2}\,[\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]$. The reflection operator represented by $[\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]$ is a group element S, and in principle, a spinor ψ also represents a group element g. But $Sg = g$ is not possible on the rotation group as S transforms a rotation into a reversal and vice versa. Although the eigenvalue equation Eq. 6 has an algebraic solution ψ, it is therefore not possible to interpret ψ as a spinor representing a group element. The problem is thus what the meaning of ψ could be. Let us therefore consider a set $\mathcal{S} = \{g_1, g_2\}$, whereby g_1 is a group element and $g_2 = Sg_1$. We have then $S(\{g_1, g_2\}) = \{g_1, g_2\}$, because S transforms g_1 into $g_2 = Sg_1$ and $g_2 = Sg_1$ into g_1. The set $\mathcal{S} = \{g_1, g_2\}$ is then a kind of generalized eigenvector of the reflection operator S. This leads to the idea that $\psi = \psi_1 + \psi_2$, whereby ψ_1 represents g_1 and $\psi_2 = [\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]\,\psi_1$ represents g_2 could represent the set $\mathcal{S} = \{g_1, g_2\}$. This is further confirmed by the explicit value for ψ when $\lambda = 1$:

$$\psi = \left(\begin{array}{c} 1 + s_z \\ s_x + \imath s_y \end{array}\right), \tag{12}$$

which corresponds to the choice $g_1 = \mathbb{1}$.[6] For a different choice of g_1, we will obtain a different quantity $\zeta = e^{\imath\chi/2}\psi$. It must be noted that this idea to associate $\psi = \psi_1 + \psi_2$ with a set can only be introduced under the form of a new definition, because a linear combination of spinors is not a new spinor, as we already explained. More generally, we could consider an ensemble of $2N$ objects O_j whereby N objects have the same orientation and parity as an object $g_1(O)$ and N objects the same orientation and parity as an object $g_2(O)$, whereby O is a reference object. When we call the numbers of objects $N_1 = N_2 = N$, the set can then be associated with $\psi = N_1\psi_1 + N_2\psi_2$. But if we normalize this quantity like a true spinor, we obtain then: $\psi = c_1\psi_1 + c_2\psi_2$, whereby $c_1 = c_2 = \frac{1}{\sqrt{2}}$ and $|c_1|^2 = |c_2|^2 = \frac{1}{2}$ represent the probabilities to find the objects in the orientations of $g_1(O)$ and $g_2(O)$. We see that this way we obtain a mixed state with exactly the interpretation given to it in quantum mechanics, viz. that a mixed state:

$$\psi = \sum_{j\in\mathcal{B}} c_j\psi_j, \quad \text{where:} \quad \sum_{j\in\mathcal{B}} |c_j|^2 = 1, \tag{13}$$

where $\mathcal{B}$ is a set, and where the sums must be replaced by integrals if the set is not countable, represents a statistical ensemble whereby the particles can be in one of the states ψ_j. The probability for this to happen is $|c_j|^2$. We can consider this as an *a posteriori* justification of the superposition principle for the wave function solutions of a Dirac or Schrödinger equation. We must note in anticipation that this rule, which is completely intuitive, will force us to consider the wave function ψ in situations where the probabilities are not summed according to the:

[6] The case of the eigenvalue $\lambda = -1$ is analogous. One just must make a different choice $\psi_2 = -[\,\mathbf{s}{\cdot}\boldsymbol{\sigma}\,]\,\psi_1$ for ψ_2.

$$\textit{incoherent rule:}\quad p = \sum_{j\in\mathcal{B}} |c_j|^2\,|\psi_j|^2, \tag{14}$$

but according to the much less intuitive:

$$\textit{coherent rule:}\quad p = |\sum_{j\in\mathcal{B}} c_j\psi_j\,|^2, \tag{15}$$

as not being obtained by the superposition principle. We will argue that the wave function is then not constructed according to a superposition principle but according to a Huyghens' principle. The motivation for defining such a distinction is a concern about mathematical rigor and clarity. There can be only one probability rule for one construction principle. There are then two different words for two different concepts, where using a same single word for two different concepts can lead to confusion, especially if the difference between the concepts is subtle enough to escape attention.

We have discovered simultaneously in this section how the necessary introduction of the mixed states replaces the one-particle description of an electron given by the Dirac-like equation by a description of statistical ensembles of electrons given by the Dirac equation. This issue remains hidden in the traditional presentation of the Dirac equation, which renders it very hard to imagine what could be going on behind the scenes of this equation. One almost has the impression that Dirac must have received the idea for his stroke of genius in a phone call from God. This impenetrability in turn leaves the door open for speculations leading to the conviction that the Dirac equation could describe a single electron. We may note that the *a posteriori* justification of the superposition principle solves also the paradox of Schrödinger's cat. The cat is not in a superposition state where it would be half dead and half alive, because the wave equation does not describe a single cat. The wave function describes a statistical ensemble of cats, whereby half of the cats are dead and half of them alive.

3 The second transgression: From orbits to orbitals

We will argue below that the second transgression discussed above, boils down to a further extension of the sets considered. Instead of the history of particles on one set $\mathcal{P}$ we will consider histories of particles on many alternative sets $\mathcal{P}'$, such that we obtain something that resembles Feynman's all-histories formulation of quantum mechanics. Instead of an all-histories approach, it will be rather a many-consistent-histories approach [6]. In fact, what we do for the case of the free-space Dirac equation in order to extrapolate the equation to $\mathbb{R}^4$ is to translate the straight line $\mathcal{P}$ over all vectors $\mathbf{w} \in \mathbb{R}^3$. This corresponds to substituting $\mathbf{r}_0 \,|\, \mathbf{r}_0 + \mathbf{w}$. The Dirac equation is then stipulated to be true on the sets $\mathcal{P}_{\mathbf{w}} = \mathbf{w} + \mathcal{P}$. Of course the equation becomes then valid over $\cup_{\mathbf{w}\in\mathbb{R}^3} \mathcal{P}_{\mathbf{w}} = \mathbb{R}^4$. In Eq. 4 we can then replace $\nabla_{\|}$ by ∇ because by construction $\nabla_{\perp}\psi_* = 0$. Here ψ_* is the extrapolated wave function. For the original wave function ψ on $\mathcal{P}$, the expression $\nabla_{\perp}\psi$ was not defined. We could have tried to define it, after introducing the *Ansatz* $\psi(\mathbf{r},t) = 0, \forall(r,t) \notin \mathcal{P}$ but this would then have led to singularities, because the result would not have been a smooth function. For this reason, the extrapolation ψ_* cannot be considered as the superposition of the wave functions $\psi_{\mathbf{w}}$. We have $\psi_* = \sum_{\mathbf{w}\in\mathbb{R}^3} \psi_{\mathbf{w}}$ but the wave functions $\psi_{\mathbf{w}}$ are solutions of mutually different wave equations defined over different sets $\mathcal{P}_{\mathbf{w}}$, and these wave functions $\psi_{\mathbf{w}}$ are also not solutions of the wave equation for ψ_* defined over $\mathbb{R}^4$.

For the free-space Dirac equation, the meaning of the extrapolation is nevertheless extremely simple. The histories become possible alternative histories over $\mathbf{w} + \mathcal{P}$ and the equation describes potentialities. The spinor $\psi(\mathbf{r},t)$ describes the motion the spinning electron would display, if it were in $(\mathbf{r},t)$. It is thus normal that also this procedure corresponds to introducing probabilities. However, we show in reference [1] (on p. 204) that we introduce this way also unexpected solutions.[7] The reason why the algebra permits to carry out the extrapolation of the definition domain unwittingly is that we only have to perform the time part of a Lorentz transformation, because the Rodrigues equation Eqs. 2-3 is purely temporal and does not depend on the position in space of the electron. This time part of the Lorentz transformation does not reflect the full information about the Lorentz transformation.

To obtain an equation that provides the same kind of description for an electron moving within a potential, we can introduce the minimal substitution, as explained in reference [1]. We can do this before or after the extrapolation procedure. When we do it before the extrapolation, the classical Dirac equation will be defined on a classical orbit, and the meaning of the extrapolation will not be as clear. The non-triviality of the procedure transpires here in several issues. The extrapolation replaces the classical orbit by a quantum mechanical orbital. The extrapolation can also entail self-consistence conditions that lead to quantization as explained in Chapter 8 of reference [1]. In the case of tunneling, the extrapolation to regions of space that are classically forbidden also gives rise to conceptual difficulties. The answer to the question what the extrapolation means will vary from case to case as we already pointed out in the Introduction. The reason for this is that different cases will involve different mechanisms.[8] This is

[7] It can be shown that the free-space Dirac equation allows also for circular orbits, because it is only based on the temporal part of the Lorentz transformation.

[8] In reference [1] we have tried to give a tentative explanation for the meaning of the extrapolation of the definition domain of the Dirac equation in the case of the calculation of the energy spectrum of the hydrogen atom. The argument runs over pages and pages (pp. 206-247). But understanding the double-slit experiment and tunneling proved to be an even much harder problem, which is why we managed to address these problems only recently.

not surprising because the solution of a wave equation is largely dictated by the symmetry of the equation and because symmetry arguments do in general not permit to nail down the underlying mechanism. This is due to the circumstance that there can exist several wildly different mechanisms that all share the same symmetry, e.g. the Bloch wave eigenfunctions $e^{\imath\mathbf{k}\cdot\mathbf{r}}$ used for phonon propagation and for diffusion in a crystal are both defined by the translational symmetry of the crystal lattice, but the mechanisms involved in these two processes are entirely different (see reference [1], pp. 32-40, p. 46, pp. 337-340). The problems become less obvious when we introduce the minimal substitution after the extrapolation, but this can only mean that we are sweeping then the difficulties surreptitiously under the carpet, because the end result of both approaches is the same. It must also be pointed out that the minimal substitution is not rigorously exact, because it fails to account for Thomas precession as discussed in [2].

4 Conventions and description of the double-slit experiment

4.1 Classical and quantum mechanical, one slit or two slits

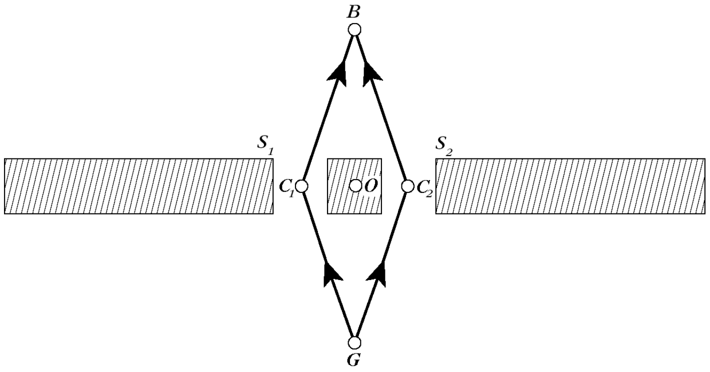

Fig. 1. Drawing of the double slit experiment, showing the notations used in the text.

The paradox of the double-slit experiment has been described in a very detailed way by Feynman in his famous lectures [4]. Let us point out the main features, assuming we are performing a double-slit experiment with electrons. In Fig. 1 we show the set-up of this experiment and summarize the meaning of the various symbols we will use. We call the slits S_1 and S_2, and their equal widths $w_1 = w_2 = w$. The distance between the centres C_1 and C_2 is called $D > w$. The centre of symmetry O is defined by $C_1O = OC_2$. We can define a reference frame $Oxyz$ whereby the x-axis coincides with the line C_1C_2, while the y-axis is perpendicular to the plane that contains the slits. The z-axis is perpendicular to the figure and parallel to the extension of the slits outside the plane. The coordinates of C_1 are thus $(-D/2, 0, 0)$, those of C_2 are $(D/2, 0, 0)$. We imagine the source G of the electrons positioned at some point $(0, d_G, 0)$ on the y-axis, with $d_G \ll 0$. We can actually put symbolically $d_G = -\infty$. In Fig. 1 the scales used for the x-direction and for the y-direction are thus completely different. With $d_G = -\infty$ the wave we would use to describe the wave function before the slits could thus be considered as a plane wave propagating along the y-axis.[9] There is a planar detector screen for the electrons far beyond the slits. The detector plane is parallel to the Oxz plane. The point B in the figure can be considered to be in the plane of this detector. When only one slit S_j, $j \in \{1, 2\}$ in the experiment is open, we will assume the source is positioned at $(-D/2, d_G, 0)$ or $(D/2, d_G, 0)$ rather than $(0, d_G, 0)$, where again $d_G = -\infty$. We will call the wavelength of the electrons λ. Let us first consider the case that only one slit S_j, $j \in \{1, 2\}$ is open. We can then consider two cases:

[9] In traditional quantum mechanics this would be interpreted literally as a plane *matter wave* ψ. Here we consider the wave function as the description of a spinning motion, which has been defined on a single path and then extrapolated to the whole Oxy-plane. We could think that the original path takes the electron through slit S_1 and then has been extrapolated to $\mathbb{R}^2$, yielding a wave function ψ_1, which is a pure state. One could argue that there is a second pure state ψ_2 obtained by symmetry $x \to -x$ from ψ_1. The true wave function $\psi = \psi_1 + \psi_2$ is then a mixed state. However, within certain limits the phase $e^{-\frac{\imath}{\hbar}(Et-\mathbf{p}_1\cdot\mathbf{r}_1)}$ of the plane wave allows for values $\mathbf{p}_2\cdot\mathbf{r}_2 = \mathbf{p}_1\cdot\mathbf{r}_1$, such that the path that takes the electron through S_2 is a history that is compatible with ψ_1. We could thus consider a description of the physics that uses only ψ_1. However, such a description of the two compatible histories is not symmetrical while the experimental set-up is. In fact, the path through S_1 has a compatibility with ψ_1 that is superior to the compatibility of the path through S_2. Such a superiority is not present in the experimental set-up. We therefore risk to introduce some bias by using just ψ_1. It is therefore better to take a wave ψ that propagates along the y-axis with a phase $e^{-\frac{\imath}{\hbar}(Et-\mathbf{p}\cdot\mathbf{r})}$ to describe the history before the slits and to consider the two actual electron paths (through S_1 or S_2) as compatible histories. We do then not consider ψ as a superposition state. Alternatively, one can introduce the mixed state $\psi = \psi_1 + \psi_2$ in order to eliminate the bias, but we will see that this is *not exact*.

- (C1) When the energy of the electrons is high enough, such that $w \gg \lambda$, we will be in the classical regime and the form of the distribution $p^{(incoh)}(\mathbf{r})$ of the impacts of the electrons on the detector screen we will measure will be similar (on a smaller length scale) to what what we would measure in a macroscopic experiment whereby we try to send tennis balls through a hole in a brick wall. This is the classical case where the probability distribution for the impacts of the electrons displays "particle behaviour" and just corresponds to the distribution for tennis balls without spin. The corresponding wave function will then be $\psi_j^{(incoh)}(\mathbf{r})$ and the probability distribution will be $p_j^{(incoh)}(\mathbf{r}) = |\psi_j^{(incoh)}(\mathbf{r})|^2$.

- (Q1) But if the energy of the electrons is low enough, such that such that $w \lesssim \lambda$, we will be in the quantum regime and the distribution will also show diffraction fringes. In this case the probability distribution for the impacts of the electrons displays "wave behaviour". The corresponding wave function will then be $\psi_j^{(coh)}(\mathbf{r})$ and the probability distribution will be $p_j^{(coh)}(\mathbf{r}) = |\psi_j^{(coh)}(\mathbf{r})|^2$.

Let us now also consider the case when both slits are open. Also here we will consider two cases, because the double-slit experiment must be set up with $w < D \lesssim \lambda$:

- (C2) When the energy of the electrons is high enough, $D \gg \lambda$, we will be in a situation that is the analog of a classical-physics experiment with tennis balls, when there are two identical holes at the same height in the brick wall. We will obtain a distribution of impacts that is the superposition $\frac{1}{2}p_1^{(incoh)} + \frac{1}{2}p_2^{(incoh)}$ of the individual probability distributions, where the normalization factor $\frac{1}{2}$ is introduced to take into account that there are now twice as much impacts, because there are two slits.[10] We could define a wave function that is a *mixed state* $\psi^{(incoh)} = c_1\psi_1^{(incoh)} + c_2\psi_2^{(incoh)}$ for this experiment, whereby $c_1 = c_2 = \frac{1}{\sqrt{2}}$. As explained in Section 2, quantum mechanics tells us that the probability that the electron is in state $\psi_j^{(incoh)}$ is given by $|c_j|^2 = \frac{1}{2}$ and that the distribution of impacts is obtained by summing the probability amplitudes *incoherently* according to the rule $p = |c_1\psi_1^{(incoh)}|^2 + |c_2\psi_2^{(incoh)}|^2$. This is the classical case where the probability distributions for the electrons (and the tennis balls) display particle behaviour.

- (Q2) When the energy of the electrons is low enough, $D \lesssim \lambda$, the probability distribution of the electron impacts on the detector screen will show interference fringes, analogous to what we observe in a macroscopic experiment with waves, e.g. macroscopic waves in a water tank wherein the waves can propagate through two openings in a wall in the middle of the tank.[11] Most importantly, the probability distribution observed is different from the sum of the probability distributions obtained in the two single-slit experiments. Text books generally claim that the wave function is now $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$, whereby $c_1 = c_2 = \frac{1}{2}$ (or $c_1 = c_2 = \frac{1}{\sqrt{2}}$ after normalization). However, this is not rigorous and in general *not exact*. It can under certain circumstances be a good approximation, but the scope of validity of the *Ansatz* is never discussed. What is always correct however, is that the probabilities are now given by $|\psi^{(coh)}|^2$. This implies that when the approximation $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ is valid, then the probability amplitudes have to be summed *coherently* according to the rule $p^{(coh)} = |c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}|^2$, which implies that we no longer sum probabilities but probability amplitudes. Here one must also take $c_1 = c_2 = \frac{1}{\sqrt{2}}$ to obtain normalization.

4.2 Caveat

The fact that text books present $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ as an *exact rule* is disturbing, because it serves as a starting basis for introducing false concepts. We have anticipated this in Section 2 by introducing the distinction between the superposition principle and Huyghens' principle. In fact, $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ seems to embody the superposition principle, but if the result is not exact, then it is not obtained from the true superposition principle. It is also difficult to reconcile that we could explain the rules $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ and $p = |\psi^{(coh)}|^2$ by using the superposition principle and simultaneously explain the rules $\psi^{(incoh)} = c_1\psi_1^{(incoh)} + c_2\psi_2^{(incoh)}$ and the corresponding probabilities $p = |c_1\psi_1^{(incoh)}|^2 + |c_2\psi_2^{(incoh)}|^2$ by using the superposition principle. We cannot credibly justify a rule that stipulates that we should sum coherently on Mondays, Wednesdays and Fridays and incoherently on the other days of the week.

As pointed out in the Introduction, the superposition principle is not justified by the group theory, because spinors belong to a curved manifold rather than to a vector space. We have been able to rationalize the superposition principle and give a precise meaning to it in terms of sets (i.e. statistical ensembles) in the case of Pauli's definition of spin (see Section 2). This rationalization leads to the rule $p = |c_1\psi_1^{(incoh)}|^2 + |c_2\psi_2^{(incoh)}|^2$. The rules $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ and $p = |\psi^{(coh)}|^2$ must therefore be wrong in principle (in the sense that they are not truly based on the superposition principle). The probabilities $p = |\psi^{(coh)}|^2$ exhibit then

[10] When incidentally $w \approx \lambda$ (such that we are in the case Q1C2) the superposition $\frac{1}{2}p_1^{(incoh)} + \frac{1}{2}p_2^{(incoh)}$ will exhibit some diffraction fringes.

[11] This is of course only true when $w \lesssim \lambda$. If incidentally $w \gg \lambda$ such that we are in the case C1Q2, the pattern will look more like a diffraction pattern.

perhaps some oscillatory behaviour, but different from the one that could be deduced in a simple way by an exact calculation from the superposition principle. A rigorous basis for explaining the oscillatory behaviour can thus not rely on the superposition principle.

In [1] we ran into this problem. We found a very different rationale to explain the oscillatory behaviour in the double-slit experiment but were unable to prove the rule $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ given by the superposition principle. The point is that we could prove ([1], pp. 330-333) that $\psi = \sum_{j\in\mathbb{Z}} c_j\chi_j$, whereby χ_j is a wave function for which the phases built up over the paths GC_1B and GC_2B are different by an angle $2\pi j$ and the different pure states χ_j do not overlap, whereby we mean that $(\forall\mathbf{r}\in\mathbb{R}^3)(\chi_j(\mathbf{r})\chi_k(\mathbf{r}) = \chi_j^2(\mathbf{r})\ \delta_{jk})$.[12] The calculus for the probabilities yields then $|\psi|^2 = \sum_j |c_j|^2|\chi_j|^2$, whereby there is no difference between the coherent and the incoherent calculation of the probabilities. But the textbook rule $\psi^{(coh)} = c_1\psi_1^{(coh)}+c_2\psi_2^{(coh)}$ cannot be justified by such a classical argument, because $\psi_1^{(coh)}$ and $\psi_2^{(coh)}$ do overlap. The coherent calculation would yield then interference fringes while the incoherent calculation would not. The two calculations are thus not equivalent when we apply them to the sum $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$, such that the incoherent calculation could then no longer be used to justify the coherent one.

We must and will justify the coherent rule in a completely different way. We will claim that it is just the rule $p = |\psi|^2$ applied to a single wave function. The rule $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ is not exact as $\psi^{(coh)}$ is not obtained by a superposition principle in the sense we have used in Section 2. In Section 2 we have used the superposition principle for solutions of a same wave equation. The rule $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ is all together different as it combines solutions from two different wave equations (corresponding to two different single-slit potentials) to propose a solution for a third wave equation (corresponding to a double-slit potential).[13] The falsification of presenting the superposition principle as an absolute truth can badly mislead somebody who is interested in the foundations of physics underlying the double-slit experiment. It could side-track him in trying to derive $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ as an exact result, while it is not. He will then run in circles for ever because the proof he tries to find simply does not exist. It would be like trying finding a rigorous proof for a wrong theorem. As long as you stay rigorous and do not make an error, you will never find such a proof, because it simply does not exist.

4.3 Summary

In summary, in certain double-slit experimental set-ups we must add probability amplitudes quadratically or "incoherently". We first calculate the squares and then take the sum. The observations are then classical and correspond to "particle behaviour". In other experimental set-ups, textbooks tell us that we must add the probability amplitudes linearly or "coherently". We first take the sum and then square the result. The observations are then quantum mechanical and correspond to "wave behaviour". However, the latter result is only qualitatively correct and not exact.

Feynman points out how classical and quantum mechanical behaviour can be observed e.g. for neutrons, which can undergo both coherent and incoherent scattering. That an electron can display both particle and wave behaviour, depending on the details of the experimental set-up is the famous "particle-wave duality", which Feynman called the only mystery of quantum mechanics. What is really mysterious is that when we are in case (Q1), then the distribution observed in case (Q2) is different from what one would obtain by summing the two-single slit distributions observed in case (Q1). Feynman highlighted this mystery by asking laconically how an electron that travels through slit S_1 to the detection screen can know that slit S_2 is open or otherwise.

5 Feynman's analysis: Knowing or not knowing which way the electron has traveled

Feynman further explained that the crucial point that determines wether the probabilities display particle or wave behaviour in a configuration where both slits are open is if the experimental set-up permits us to find out through which one of the two slits the electron has travelled. Of course we can know this, when one of the slits is closed, but as here we are discussing the case where

[12] There is an alternative logical option, which is to consider that there is only one single wave function whose definition domain can be subdivided in patches $\mathcal{S}_j$ over which the wave function is different from zero and the phases built up over the paths GC_1B and GC_2B are different by an angle $2\pi j$. In both options the phase difference cannot make a jump of 2π when we change one of the paths over an infinitesimal amount. Therefore, if there is only one wave function, then these patches must be separated by regions of space where the wave function vanishes. This single wave function can then be broken up in different wave functions χ_j that are non-zero over the interior of their definition domains $\mathcal{S}_j$. When there are different wave functions χ_j, the absence of overlap between the patches seems to be no longer a compelling requisite, as the probabilities must then anyway necessarily be calculated according to $|\psi|^2 = \sum_j |c_j|^2|\chi_j|^2$. However, overlap can also be excluded here based on the fact that on the patches the electron is traveling in free space. From a classical viewpoint (which we adopt in the whole paper) the result within a putative overlap between two patches should therefore be equal because identical causes must produce identical results. But in the overlap of different patches, the results would by definition be different by a phase difference that is a non-zero multiple of 2π. This way considering that there is only one or several wave functions boils down to the same thing.

[13] Note that we encountered this situation already in the description of the extrapolation procedure in Section 3.

both slits are open, the criterion at stake here must be a different one than just knowing which slits are open. It is a criterion for quantum behaviour, that is more refined than the inequality $w < D \lesssim \lambda$ given above.

E.g. if we shine a light beam on the region of space just behind slit S_1, the interaction of the electron with a photon of the light beam may tell us that the electron has traveled through slit S_1. The probability amplitudes will then exhibit particle behaviour. In the case of neutron scattering we may also discover which trajectory a neutron has traveled when it has flipped its spin and simultaneously flipped the spin of a nucleus in the experimental set-up in a spin-spin exchange interaction. The spin flip of the nucleus is a mark left by the passage of the neutron revealing that the neutron has interacted with the given nucleus. The scattering is then incoherent. If there is no way to find out through which one of the two slits the electron has travelled, then probabilities will exhibit wave behaviour. The analogous situation in the neutron scattering experiment will be that the neutron has interacted with the nuclei in the set-up without inducing any spin flip. There is not a single trace the neutron has left of its passage. The scattering is then coherent.

An experiment whereby we put the detector screen close to the slits will also permit us to find out to a certain extent through which slit the electron will have travelled. In the region close to the slits we will thus not so much observe wave behaviour. This could be due to two reasons. The first one is that ψ_2 becomes very small close to slit S_1 and ψ_1 very small close to slit S_2. A second reason could be that the rule $\psi^{(coh)} = c_1\psi_1^{(coh)} + c_2\psi_2^{(coh)}$ does not hold in this region, and becomes more accurate at a long distance behind the slits. The idea that the rule is not exact is in agreement with the results of the work of Sinha *et al.* [7].

When an electron does not leave any trace of its passage through the device, in analogy with the case of slow-neutron scattering without any spin-flip, then its interaction with the set-up is coherent. It is then impossible in principle to know which way the electron has traveled. Even Hercule Poirot will not be able figuring it out. In the double-slit experiment, this means that not a single atom or electron in the set-up carries a mark of the passage of the electron. This can only be true if not a single atom or electron of the device has been scattered by its interaction with the electron, because the recoil produced by the scattering would create a phonon or an electronic transition within the measuring device and at least in principle this could be detected. The first excited electronic level might be too high to allow a transition or the interaction must be recoilless like in the Mössbauer effect, where it is the set-up as a whole which recoils instead of a single atom (Such a recoil of a macroscopic device is undetectable, because the devise is too massive). Or it should not induce a spin flip in a neutron scattering experiment. When the interaction of the electron with the measuring device does leave a mark in the device, e.g. in the form of a recoil, then the interaction process is incoherent. In a neutron scattering experiment, the incoherent interaction that left its mark on the device could be a spin flip rather than a recoil. However, when the interaction in the double-slit experiment is coherent, then still the probability distribution is not the superposition of the coherent single-slit probability distributions.

6 Thesis: The quantum weirdness does not reside in the properties of the particles

6.1 Two principles

This looks very mysterious indeed and seems to defy all daily-life intuition or common sense. We are used to describe this as weird quantum behaviour. The weirdness is often considered as a quantum paradox. We must point out that this paradox does not reside in the physics. The basic cause for it is not a weird property of the electron. To explain this we must develop two points, that we introduce here shortly.

(1) The first one will be amply discussed In Subsection 7.1 and is of a general nature. We will see that we can summarize it by stating that:

> *We are not measuring electrons with the experimental set-up,*
> *We are measuring the experimental set-up with electrons !*

Moreover, the complete information about the set-up we can obtain from the electrons (if we use enough of them) is non-local, because the geometry of the set-up is non-local!

(2) The second point is very specific for the double-slit experiment and must be combined with the general ideas evoked under point (1). For the double-slit experiment, we will argue in addition that the paradox is a probability paradox that comes about when we can no longer use binary logic when it comes down to answering simple yes-or-no questions, and we are forced to admit in all honesty for the possibility that the answer to a question can also be *undecidable*.[14] We will here shortly introduce the concept of undecidability, before applying it to the double-slit experiment in Subsection 7.3 and show how it connects with Heisenberg's uncertainty principle. Subsection 7.2 will explain in detail the difference we want to define between a superposition principle and a Huyghens' principle.

[14] Let us state right ahead for the reader who may feel that this is farfetched, that we will show that this expresses something he is much more familiar with, viz. Heisenberg's uncertainty principle. The undecidability does not apply to the electrons but to the set-up (see Section 12).

6.2 Undecidability

Questions that are undecidable are well known in mathematics. Examples occur e.g. in Gödel's theorem or in the question if there exists a cardinal number between the cardinal $\aleph_0$ of the set of the integers and the cardinal $2^{\aleph_0}$ of the set of real numbers [8]. Paul Cohen [9] has shown in 1963 that this question is undecidable within the Zermelo-Fraenkel set theory with the axiom of choice included.[15] The existence of such undecidable questions may look hilarious to common sense but this does not need to be. In fact, the reason for the existence of such undecidable questions is that the set of axioms of the theory is incomplete. We can complete then the theory by adding an axiom telling the answer to the question is "yes", or by adding an axiom telling the answer to the question is "no". The two alternatives lead to two different axiomatic systems and thus to two different theories. An example of this are Euclidean and hyperbolic geometry [10]. In Euclidean geometry one has added on the fifth parallels postulate to the first four postulates of Euclides, while in hyperbolic geometry one has added on an alternative postulate that is at variance with the parallels postulate. The axiom one has to add can be considered as information that was lacking in the set of axioms. Without adding it one cannot address the yes-or-no question which reveals that the axiomatic system without the parallels postulate added is incomplete. As Kurt Gödel has shown, we will almost always run eventually into such a problem of incompleteness.

When the interactions are coherent in the double-slit experiment, the question through which one of the two slits the electron has traveled is very obviously also undecidable. Just like in mathematics, this is due to lack of information. We just do not have the information that could permit us telling which way the electron has gone.[16] According to common-sense intuition this would not be too much of a problem in performing our probability calculus, as the undecidability is just experimental. We reckon that in reality, the electron must have gone through one of the slits and not through the other. We argue then that we can just assume that half of the electrons went one way, and the other half of the electrons the other way. But rigorous logic shows that the way we apply this reasoning to calculate the probabilities contains a fallacy, while by doing the logic correctly we just reproduce the quantum mechanical result! [17] In order to show this we will introduce a number of assumptions that are all intuitive.

7 Explicit formulation of all assumptions underlying the philosophy of our approach

7.1 Assumption 1: Particles behave like particles, statistical ensembles of particles behave like waves

There is no particle-wave duality. Electrons are always particles, never waves. Electrons never travel like a wave through both slits simultaneously. That electrons are always particles is evidenced by the fact that a detector detects always a full electron

[15] A whole school of thought in mathematics founded by Brouwer (the intuitionists) does not accept proofs based on a *reductio ab absurdo* due to the idea that a theorem could be undecidable. In a sense, the *reductio ab absurdo* can be considered as an untidy short-cut. The theorem you want to prove could be undecidable, which means that there does not exist a proof. At that stage, you need a supplementary axiom. You can use the affirmation of the theorem as the supplementary axiom or it's negation. But the negation leads to a contradiction. The only extension of the axioms that does not obviously contain a contradiction is then the one that has the axiom of the affirmation of the theorem added. The theory contains then a new truth, but this truth has the status of an axiom rather than the status of a theorem, because we could not prove it, in the sense that we proved that the theorem was undecidable. This way, we build up theories in the hope that we can avoid that contradictions will occur.

[16] There exists a video entitled *"Probability & uncertainty - the quantum mechanical view of nature"*, with a lecture of Feynman on the double-slit experiment. In this lecture Feynman considers three possibilities for the history of an electron: "slit 1", "slit 2", and "do not know". The third option corresponds exactly to this concept of undecidability. This is worked out with many examples in reference [4], to show that there is a one-to-one correspondence between undecidability and coherence.

[17] Two remarks are due here:

(1) As evoked in Section 2, one can derive the Dirac equation in a completely classical way from the *Ansatz* that the electron spins. The equation is then defined on a classical path $\mathcal{P}_1$. We can call it the equation E_1. However when we solve it, we search for a solution over the whole of $\mathbb{R}^4$. We extrapolate thus the definition domain of the differential equation from the path $\mathcal{P}_1$ to $\mathbb{R}^4$. Due to the difference of definition domains, we will call this extrapolated differential equation, the equation E. We might also derive an equation E_2 on a different path $\mathcal{P}_2$ and extrapolate it, and obtain the same equation E. The equation E describes then several possible histories which are mutually consistent in the sense that they can be described by the same equation E. Of course to be consistent two histories must satisfy certain conditions. We find here back the concept of consistent histories introduced by Griffiths [6]. For the equation E we can also justify that the probabilities can be obtained from the wave function ψ by using $|\psi|^2$. This is done in all textbooks by showing that one can derive a continuity equation for the probability charge-current from the Dirac equation. We can calculate this way the probabilities for the two single-slit experiments and the double-slit experiment and perfectly reproduce this way the double-slit experiment paradox within the theory. The mathematical solution of the paradox is then that slightly different boundary conditions for a differential equation can give rise to vastly different solutions (see below). The different boundary conditions (in combination with the value of the wavelength) are then the means by which we can express if a question is decidable or not decidable in the description of the experiment.

(2) It might provoke disbelief that we would have to abandon our binary logic based on the mutually exclusive concepts true and false, in order to be able to make sense of quantum mechanics and the idea might meet resistance, even though we tried to iron out some of this resistance in the preceding lines. Choice of accurate terminology can be very important. Would the resistance of the reader have been the same if we had used the word *uncertainty* instead of *undecidability*? This raises the very interesting question about the exact relation between the two concepts, on which we will dwell below.

at a time, never a fraction of an electron. It is the probability distributions of the electrons which display wave behaviour, not the electrons. In the preceding lines we have already tacitly anticipated the introduction of this postulate, by carefully phrasing our ideas to make them consistent with it. This postulate only reflects literally what quantum mechanics says, viz. that the wave function is a probability amplitude. Although this sharp dichotomy is very clearly present in the rules, we seem to find it difficult to perceive it. This is due to the fact that there has been a tendency to blur this very sharp image again by imposing a doctrine that wants to read more into this issue by adding additional interpretation. But there is absolutely no need for such additional interpretation: *In claro non interpretatur!* All one adds is over-interpretation.[18] The wave functions ψ do not describe single electrons but a probability distribution for electrons corresponding to a given set-up. Such a probability distribution is defined over the whole space accessible to the electron. It includes thus information about the whole experimental set-up. A single electron can in principle not see if the other slit is open or otherwise, because it cannot probe non-local information.[19] But the probability distributions and the wave functions contain this kind of global non-local information as they are defined simultaneously over whole space. This information is based on a simultaneous, global and therefore non-local description of all parts of the whole experimental set-up.[20] The wave function is therefore defined in a non-local way while the electron cannot have non-local interactions.

That such a non-locality of the formulation is not in contradiction with relativity can be explained by pointing out how also Lorentz frames used to write the Lorentz transformations are non-local because they assume that all clocks in the frame are synchronized up to infinite distance (see e.g. p. 278 in reference [1]). It is by no means possible to achieve this, such that the very tool of a Lorentz frame conceptually violates the theory of relativity. In general, this is without consequences for the theory, because we never use the Lorentz transformations truly over the whole of $\mathbb{R}^4$ to describe some physical problem. We can always restrict the transformation to a patch of $\mathbb{R}^4$, but we never do this. We can justify this *a posteriori* by claiming that this was in order to avoid burdening the presentation. This way we catch up for the error that we did not realize that a Lorentz frame is conceptually not an exact tool. What this discussion shows is that the very concept of a wave function $\psi(\mathbf{r}, t)$ defined over whole $\mathbb{R}^3$ at an instant t is also non-local, and therefore also conceptually not an exact tool. It is not the physics here that is non-local. It is our tool we use to describe the physics which is. We find non-locality back in the description of the double-slit experiment in Bohm's interpretation of quantum mechanics [12]. Non-locality is thus *a priori* not a physical issue as one might be tempted to

[18] Perhaps this due to the historical introduction by de Broglie of the concept of matter waves, while the wave function corresponds only to the (extrapolation to $\mathbb{R}^4$ of the) description of the spinning motion of the electron over a path $\mathcal{P}$ as explained in Section 1 (see also Footnote 9).

[19] In reality, the electron could encounter different potentials in the two-slit and single-slit experiments. As we only describe the experiments based on symmetry arguments (whereby we do not want to refer only to symmetries of Euclidean geometry but also to physical conservation laws), we elude describing the detailed underlying mechanism. This mechanism is certainly not the same for photons and electrons. For electrons it might e.g. involve polarizing the material of the slits (see Footnote 24). The induced charge distribution may then be slightly affected by the presence of a second slit. Photons on the other hand may induce oscillations of dipoles, which then in turn could emit photons. This leads to the idea that the textbook treatment of the double-slit experiment is an idealized and simplified, abstract toy model that manages to capture the essence of some qualitative features of the physics that occur as a common denominator in both the photon and the electron experiments, but not the quantitative details of these experiments. What pleads for this is that electrons, photons and even other particles (like alpha particles, neutrons and protons) have different physics and obey different wave equations. As we will see, this essence is the undecidability built in into the set-up (see Subsection 6.2 and Subsection 7.3). A quantitative exact calculation of the interference patterns might require describing the whole case-dependent mechanism. Such situations occur in other instances in physics. E.g. Ohm's law has a huge range of validity. In different parts of this range of validity there are certainly many different mechanisms responsible for it. The idealized description of the experimental set-up may even impede us to see the correct formalism. We discussed e.g. that for neutron scattering the reason why the question through which slit the neutron has gone is decidable or not decidable resides in the presence or absence of spin flip. The idealized description does not address this issue. It tells us if the question is undecidable or decidable for a wrong reason: the symmetry of the experimental set-up combined with some considerations ($D \lesssim \lambda$ and $w \lesssim \lambda$) about the wave length of the neutron. These considerations provide a good guess for the answer to the question if the scattering process will leave a mark on the system or otherwise, i.e. if it will be incoherent or coherent. They are not exact, but just a crude rule of thumb, and they have historically also always been presented as such (The Heisenberg uncertainty relations are also such a rule of thumb but they are covering actually only a special subset of the possible undecidable situations). But as the rule of thumb shortcuts the task of providing the mechanism, the result looks mysterious. The same aspects will also transpire in the description of the tunneling experiment for an electron, where the description in terms of a constant potential barrier does not allow for any description of the band structure, while this may be a fundamental ingredient to explain the process of the creation of an electron-hole pair in the solid [11]. Similarly, a potential barrier with tetrahedral symmetry might be used to describe the tunneling of a CH_4 molecule. This is absolutely clueless with respect to the possible mechanism, which for this particular case seems almost beyond imagination or guessing. Finally, without the wavelength criterion we would even not be able to distinguish between coherent and incoherent scattering when both slits are open in the double-slit experiment, because the symmetry does not provide a clue about the mechanism.

[20] Indeed, the very description itself of the set-up we use to define the wave equation does not provide any clue as to which way an electron will have traveled. It leaves all possibilities open. This lack of information must then also show up in the wave function, because it cannot contain more information than the equation that defines it. If you do not tell in the description of the set-up which way the electron travels, why would the solution of the wave equation then tell you which way it would have gone? But the idealized description fails to describe the microscopic information that might enable to tell which way the particle has gone, e.g. a spin flip for the neutron. The macroscopic description is therefore not a real theory. It just is a very fortunate shortcut to the detailed microscopic argument that a correct theory might turn out.

conclude from Bohm's result. It would require more study to check if Bohm's result implies some non-locality that goes beyond the kind of trivial non-locality we are describing here.

As we highlighted above, in the double-slit experiment we are not measuring electrons with the experimental set-up. We are measuring the set-up with electrons. These electrons will measure the global information about the set-up and it is therefore that your wave function must be global. The wave function must be able to account for every detail of the set-up because the electrons will explore every detail of the set-up provided you are sending out enough of them to make the full exploration. It is also therefore that you must extrapolate the wave function to whole space-time. That is the only way to be sure that it will reflect the full global information.

We can render these ideas clear by an analogy. Imagine a country that sends out spies to an enemy country. The electrons behave as this army of spies. The double-slit set-up is the enemy country. The physicist is the country that sends out the spies. Each spy is sent to a different part of the enemy's country, chosen by a random generator. They will all take photographs of the part of the enemy country they end up in. The spies may have an action radius of only a kilometer. Some of the photographs of different spies will overlap. These photographs correspond to the spots left by the electrons on your detector. If the army of spies you send out is large enough, then in the end the army will have made enough photographs to assemble a very detailed complete map of the country. That map corresponds to the interference pattern. We will argue later on in Subsection 9.1 that this interference pattern presents the information about the experimental set-up. It does not present this information directly but in an equivalent way, by a Fourier transform. The spies are not correlated, but the information about the country is correlated, it is the information you put on a map. You will see straight lines (roads). None of your spies has seen the global picture. None of them will have seen that long straight road that stretches out for thousands of miles (and is a kind of correlation). They may just have seen a kilometer of it. They have only seen the local picture. The global picture, the global information about the enemy country is non-local, and contains correlations, but it can nevertheless be obtained if you send out enough spies to explore the whole country, and it will show on the map assembled. That is what we are aiming at by invoking the non-locality of the Lorentz frame and the non-locality of the wave function. Each electron sees at the best one slit. Well, it even does not see a slit, it is just other electrons, which you do not measure, that are killed when they hit a part of the set-up. The wave function contains also information about electrons that you may not measure. And that is a point you will never be able to make sense of if you consider the wave function as information about the electrons you measure. This point could e.g. intervene in delayed-choice experiments [13] (see Footnote 30). But the global information gathered by many electrons contains the information how many slits are open. Because for each set-up that global information is the complete global information about your set-up contained in the wave function. You need many single electrons to collect that global information. You may (e.g. in the double-slit experiment) need also many electrons that may just not find their way to your detector. *The global wave function contains also information about the electrons you have not detected, because they are e.g. reflected by the set-up and end up in the reflected wave, or absorbed (which is an incoherent process),* and this defines a part of the values of the transmission function C defined in Eq. 16 below. And as far as those electrons that are being detected are concerned, a single one just gives you one impact on the detector screen. That is almost no information. Such an impact is a Dirac delta measure, which is the Fourier transform of a flat distribution. It contains hardly any information about the set-up because it does not provide any contrast.[21]

The description of the experimental set-up that we use to calculate a wave function is conventionally highly idealized and simplified. Writing an equation that would make it possible to take into account all atoms of the macroscopic device in the experimental set-up is a hopeless task. Moreover, the total number of atoms in "identical" experimental set-ups is only approximately identical. What we present is always a figure like Fig.1, and we could write down a Schrödinger or Dirac equation based on this figure, assuming e.g. that the macroscopic set-up presents an impenetrable, infinite potential barrier to the electron. But with such a simplified *Ansatz* we can never describe the atom that could bear the mark of the passage of the electron. Therefore, solving the equations with such an infinite potential exactly can only yield the coherent wave function. For large electron energies, this wave function will oscillate furiously and reach the limit of the incoherent wave function. We must then use *ad hoc* rules to obtain the result for the incoherent case or take refuge to a different description of the same set-up. The advantage of the infinite potential is however that it introduces very simple boundary conditions, which catch the essence of the probability paradox (see below in Footnote 23), and make it easy to carry out and understand the extrapolation of the wave function from $\mathcal{P}$ to $\mathbb{R}^4$. It is rather logical to assume that it is the probability distribution which "knows" if the other slit was open or otherwise. As said an electron cannot sense non-local information. Furthermore, a single electron impact on a detector plane does not tell us very much about the detailed probability distribution. Only by making the statistics of many impacts can we collect knowledge about the detailed probability distribution.[22]

[21] A very nice illustration of how the information gradually builds up is given in the work of Tonomura *et al.* [14]

[22] The critical reader may object that some photon correlation experiments refute the validity of our attempts to make sense of quantum mechanics based on a completely classical philosophy because they violate some Bell-type inequalities. But the Bell-type inequalities are based on the assumption that there exists a unique *common hidden-variable distribution* that can be used *in all the various configurations of the experimental set-up*. These configurations are e.g. the different combinations of polarizer settings. The inequality is derived from a simpler inequality by integrating over this common distribution. The idea is that from e.g. $a + b > 0$ it must follow that $a\rho(Q)dQ + b\rho(Q)dQ > 0$, when $\rho(Q) > 0, \forall Q$. However in [1] we show that under certain circumstances this assumption of a unique common distribution is not tenable. The quantities in the simpler inequality must then be integrated over different distributions in order to obtain the expressions for the measured quantities and the Bell-type inequality can then no longer be derived since from $a + b > 0$, it does not follow that $a\rho_1(Q)dQ + b\rho_2(Q)dQ > 0$.

7.2 Assumption 2: There is a fundamental difference between the superposition principle and Huyghens' principle

The good question is thus not if the electron sees if the other slit is open or otherwise. That is a leading question. The electron never does. A good question is how the electron interacts with the measuring device, because even when both slits are open we can observe cases (Q2) and (C2), depending on the electron energy. Similarly, when only a single slit is open, we can observe cases (Q1) and (C1). What tells the cases (Q2) and (C2), or (Q1) and (C1) apart is the question wether the interaction was coherent or incoherent. When the interaction was coherent, it has become impossible in principle to know which way the electron has traveled. It is the full distribution that knows if both slits are open or otherwise, because it can be calculated from a wave function that is constructed taking into account wether both slits are open or otherwise, using a probability rule that will imply if the processes were coherent or incoherent. The mystery does thus not reside within the electron but in the set-up of the experiment which acts as a filter for wave functions. If we make the idealization that the material of the set-up presents an infinite potential to the electrons, then the wave function must vanish at the boundaries of the set-up. These conditions are boundary conditions for the wave function. By taking into account the boundary conditions in the Schrödinger equation, the information will be rigorously taken into account and make its way to the final solution, such that it represents exactly the solution that represents the information available.[23]

The two single-slit experiments and the double-slit experiment result then in three different boundary conditions. As different boundary conditions can lead to wildly different solutions, the superposition principle is *a priori* not justified. Sometimes the linearity of the Schrödinger and Dirac equations is invoked to justify the superposition principle, but this can only be true for solutions of the equation that share the same boundary conditions. What we do in the double-slit experiment is adding the solution ψ_1 for the Schrödinger equation with a potential V_1 to the solution ψ_2 of the Schrödinger equation with a completely different potential V_2 and claim that the solution ψ_3 of the Schrödinger equation with yet another completely different potential V_3 is given by $\psi_3 = \psi_1 + \psi_2$. Such a procedure cannot at all be justified by the linearity of the Schrödinger equation with the potential V_3. The procedure that tells us that we should take $\psi_3 = \psi_1 + \psi_2$ is thus in lack of proof. If we cal $\mathcal{S}_j$ the set of points where $V_j(\mathbf{r}) = 0$, then $\mathcal{S}_3 = \mathcal{S}_1 \cup \mathcal{S}_2$, and over $\mathcal{S}_1 \cap \mathcal{S}_2$ the wave equations are identical. However, the other points define supplementary boundary conditions and these are different for V_1, V_2 and V_3. Therefore the procedure $\psi_3 = \psi_1 + \psi_2$ is at least *in principle* wrong. One cannot always represent the effect of two causes as their sum.[24] But from now on we will assume that this rule could be a good approximation to the exact result. We will see that the justification for this is not a superposition principle but a Huyghens' principle. We think it is important to make a very clear distinction between these two principles. The superposition principle applies to two solutions of a same equation. Huyghens' principle applies within a single wave function of an equation. The idea is that the Huyghens' principle could become instrumental when we are able to meet certain boundary conditions for the wave function on a patch $\mathcal{S}_1$, without considering those on a different patch $\mathcal{S}_2$, and vice versa, while it may still not at all be obvious how to meet both boundary conditions simultaneously together on $\mathcal{S}_1 \cup \mathcal{S}_2$ when $\mathcal{S}_1 \cap \mathcal{S}_2 \neq \emptyset$.

The distinction between Huyghens' principle and the superposition principle is as follows. It is a difference between coherent and incoherent histories. When we extrapolate the wave equation to whole $\mathbb{R}^4$ we can only keep coherent (i.e. mutually consistent) histories in a single wave. If two histories cannot be combined consistently into a single wave-function we must collect them into different wave functions. We have then several wave functions to which we can apply the superposition principle. The meaning to be given to a linear combination has been explained in Section 2: The linear combination $\sum_{j=1}^{n} c_j\psi_j$ represents a set of spinors containing $N\,|c_j|^2|$ spinors of the type ψ_j. Each spinor represents a possible state. These states are pure states and

The fact that there does not always exist a common probability distribution is well known in mathematics as Gleason's theorem [15]. It is curious to notice that certain physicists [16] consider this a confirmation of the claims made on the basis of Aspect's experiments [17], while in reality it refutes them because it shows that the Bell inequalities are wrong in the sense that they cannot be considered as an expression that would be universally valid for all local hidden-variables theories.

[23] In reference [1] we have derived the Dirac equation from scratch for a free electron. We can derive the Dirac equation for an electron in a potential from this by a substitution. The minimal substitution is not completely correct as it neglects e.g. the Thomas precession. But in the present context we assume that the equation with the minimal substitution is correct. The idea is then that the Dirac equation describes correctly the physical situation both in the single-slit and in the double-slit experiment. The paradox that intuitively the single-slit and the double-slit solutions seem to be incompatible is then just the paradox that different boundary conditions can lead to vastly different solutions. We encounter such a sensitivity to boundary conditions in the Dirichlet problem. The Dirichlet problem is the problem of finding a function which is the solution of a partial differential equation over the interior of a given region and satisfies the condition of taking prescribed values on the boundary of the region. The Schrödinger equations whereby we consider the potential barrier of the single-slit and double-slit experiments to be infinite correspond to such Dirichlet problems. As pointed out, these boundary conditions for the double-slit potential express in combination with the wavelength if the question through which slit the electron has traveled is decidable or otherwise.

[24] A good illustration of this could be the following. An electron approaching a slit S_j could attract holes and repulse electrons in the material surrounding the slit. The resulting probability distribution of the holes would have a behaviour which resembles L_jC_j, where L_j would be a Lorentzian centered at the position of the electron in the slit S_j and C_j a function that is zero over the slit S_j and one elsewhere. The true function could be different from a Lorentzian and what it really is does not matter. The idea is here only to describe the increasing or decreasing behaviour of the probability distribution of the holes. This probability distribution could have an effect on the wave function. Bluntly applying a sum rule for the probabilities or the probability amplitudes of the holes in the double-slit configuration would yield $L_1C_1 + L_2C_2$. This function would still contain a reminiscence of the two local maxima of L_1 and L_2, which is certainly wrong. The result should only display the reminiscence of one local maximum of L_j, because an electron approaches only one slit S_j. The correct result should be $L_jC_1C_2 \neq L_1C_1 + L_2C_2$.

mutually orthogonal according to some criterion. Such a criterion could be e.g. that $\int_{\mathbb{R}^3} [\,\psi_j^*(\mathbf{r})\psi_k(\mathbf{r}) + \psi_k^*(\mathbf{r})\psi_j(\mathbf{r})\,]\, d\mathbf{r} = 0$, which leaves open the possibility that there could be values $\mathbf{r} \in \mathbb{R}^3$ for which $\psi_j(\mathbf{r})\psi_k(\mathbf{r}) \neq 0$.[25] After normalizing we can then state that this set of spinors represents a statistical ensemble where a fraction $|c_j|^2|$ of the particles is in the state ψ_j. The probability to find the particle in this state is then $|c_j|^2$, which is what the quantum rule states. In [1] and Section 2 we show that we become forced to introduce this rule to make sense of Pauli's formalism of spin, and also to take the ultimate step in a rigorous, *deductive* derivation of the Dirac equation. This is evidence for the fact that the wave functions in the Pauli and Dirac equations are not pure spinors, but special linear combinations of them from the group ring. They are special because not all linear combinations are allowed, just one. Such linear combinations represent thus statistical ensembles as we just explained.

But in the case of the double-slit experiment the rule $\psi_3 = \psi_1 + \psi_2$ is then still not justified, because we are combining solutions from different equations. It would therefore be of interest to note such sums differently, e.g. under the form $\psi_3 = \psi_1 \boxplus \psi_2$, or $\boxplus_{j=1}^{2} c_j\psi_j$, to indicate that ψ_1 and ψ_2 are not solutions of the equation with potential V_3 and that the result $\psi_3 = \psi_1 \boxplus \psi_2$ is not exact but only a good approximation. In anticipation of the forthcoming discussion, we could then say that $\psi = \psi_1 \boxplus \psi_2$ indicates that the sum is obtained from a Huyghens' principle while $\psi = c_1\psi_1 + c_2\psi_2$ is obtained from a true superposition principle. We are using Huyghen's principle to construct a single wave function that contains only coherent (i.e. consistent) histories. We may note that the mere fact that $\psi = \psi_1 \boxplus \psi_2$ is not exact, already points out that there is something wrong with our brute-force ideas about summing, not only for probabilities, but even for probability amplitudes.

7.3 Assumption 3: The double-slit experiment involves also undecidability (uncertainty)

The double-slit paradox is a probability paradox. It is a paradox about how we calculate probabilities for parameters whose values are undecidable. When we state that we cannot possibly know through which one of the two slits the electron has traveled, we must take into account this piece of information self-consistently. The answer to the question if the electron has gone through slit S_1 is then not "yes" or "no" but "undecidable". *Mutatis mutandis*, the same requirement of self-consistence applies when we do know through which one of the two slits the electron has traveled. We are not used to undecidable questions in daily-life experience and it looks tantalizing that such questions could exist. But as we pointed out, they do occur in mathematics.

In any case, we cannot assume at one stage of the argument that we cannot know for any of the histories through which slit the electron in the particular history has gone because it has not left the slightest mark on the measuring device and act in another stage of the argument as though we know through which slit the electron has gone. The fallacy in our reasoning could be that an ensemble $\mathcal{H}^{(coh)}$ of undecidable histories $h_\mu^{(coh)}$ of recoilless passages through slits S_1 or S_2 cannot be obtained by making the conjunction $\mathcal{H}_1^{(coh)} \cup \mathcal{H}_2^{(coh)}$ of two ensembles $\mathcal{H}_1^{(coh)}$, and $\mathcal{H}_2^{(coh)}$ of "decided histories" $h_{j,\nu}^{(coh)}$ of recoilless passages through slit S_1 and recoilless passages through slit S_2. Such "decided histories" $h_{j,\nu}^{(coh)}$ just do not exist. How the hell would we assign a single element $h_\mu^{(coh)} \in \mathcal{H}^{(coh)}$ to $\mathcal{H}_1^{(coh)}$ or $\mathcal{H}_2^{(coh)}$? We have no objective criterion that entitles us to perform such an operation and our choices would be completely arbitrary. The set $\mathcal{H}^{(coh)}$ cannot be split objectively into two subsets.

It may look surprising and non-intuitive that this fact has significant consequences. The surprise resides in the fact that we can use arguments of statistical mechanics. Against all odds, we attribute an arbitrary label to each "undecided history" in $\mathcal{H}^{(coh)}$. Of course, we are then lying, because we act as though we would know through which slit the particle has gone, while we do not. But we argue that we do this only to take into account that the history must have taken the particle either through S_1 or through S_2. We must then somewhere make up for our lie, and to do this we consider all possible ways to perform this labeling in order to calculate a statistical average. This procedure will then take into account that in reality we did not know. But the factual truth shows that this procedure does not reproduce the experimental results. From an abstract and rigorous logical viewpoint we should not be surprised by this outcome. It is perfectly rigorous and logical. To discuss this we must consider two cases. The problem is complicated by the fact that it is not only a matter of which slits are open or closed.

(1) To be decided, a history must contain an incoherent interaction with the apparatus. It is just not true that we could obtain the ensemble of undecidable histories $\mathcal{H}^{(coh)}$ as the union of two ensembles of histories $\mathcal{H}_1^{(incoh)}$, and $\mathcal{H}_2^{(incoh)}$ in which the electron has interacted creating a recoil in the set-up: $\mathcal{H}^{(coh)} \neq \mathcal{H}^{(incoh)} = \mathcal{H}_1^{(incoh)} \cup \mathcal{H}_2^{(incoh)}$ (Here the indices j refer to the slit S_j through which the electron has gone). In both sets $\mathcal{H}_j^{(incoh)}$ of histories where the interactions with the set-up are incoherent there is an atom recoiling within the set-up, while in the whole set $\mathcal{H}^{(coh)}$ of histories where the interactions with the set-up are coherent, there is not a single atom recoiling. The ensemble $\mathcal{H}^{(coh)}$ in the coherent case can thus not be constructed through a randomization procedure of taking the union of the ensembles $\mathcal{H}_j^{(incoh)}$ that occur in the incoherent case. An "undecided" randomized ensemble of decidable histories is not the same as an ensemble of undecidable histories. We contradict ourselves when we assume that the histories are decided in one stage of the argument and that they are undecided in another stage of the argument. The idea that we

[25] The circumstance that $\exists \mathbf{r} \in \mathbb{R}^3 \parallel \psi_j(\mathbf{r})\psi_k(\mathbf{r}) \neq 0$ does not lead to an interference term in the superposition principle, clearly indicates that the probability calculus according to the superposition principle is very different from the calculus we apply when we are faced with interference. As we will argue, interference is based on a Huyghens' principle, which stresses the importance of clearly distinguishing the superposition principle and Huyghens' principle.

can label the histories by the labels S_1 or S_2 occurs in a theory based on a system of axioms $\mathscr{A}_1$, while the undecided histories occur in a theory based on an all together different system of axioms $\mathscr{A}_2$. The randomization procedure is thus analogous to an argument that would blend theorems from hyperbolic geometry and from Euclidean geometry together. We have not proved a theorem that would justify the algorithm based on the randomization procedure for $\mathscr{A}_2$. We have just assumed that it would be correct in $\mathscr{A}_2$ because it is correct in $\mathscr{A}_1$. That we are using the randomization procedure reflects that we are just not able to believe that $\mathscr{A}_2$ would exist in the real world. We figure that the randomization procedure will make up for the contradiction, but this is wrong, because we have lost sight of the point that the "undecided histories" and "decided histories" are fundamentally different. They can be differentiated by inspection, by checking if the interaction has been coherent or incoherent, and must thus be described by different systems $\mathscr{A}_1$ and $\mathscr{A}_2$. This is an error we make when we take our macroscopic intuition as guidance to think about the microscopic situation.

(2) However, this reasoning falls apart when also in the single-slit experiments the interactions are coherent. We must then reason in a different way. Footnote 23 already points out that this could be just a paradox of the sensitivity to boundary conditions of the Dirichlet problem. Remember that the wave function contains also information about electrons you are not detecting. In an experiment where slit S_1 is open, it contains also information about the electrons that have "died" (in terms of transmission) on the parts of the screen where slit S_2 could have been open. In a full solution of the wave function, you would treat these as reflected or absorbed waves. The wave equation is non-local and contextual. The decided histories occur in a context that contains information about the electrons you do not detect but have "died" on other parts of the set-up. This is exactly what the different boundary conditions imply. The Dirichlet problem is difficult, because it requires satisfying the boundary conditions globally, not just locally. And even in this case we are still lying when we use the statistical procedure because we act as though the particle that has gone through a slit S_j in the double-slit experiment would be described by an equation with the boundary values for a single-slit experiment. In other words, in the statistical procedure we ignore the warning we wanted to issue by introducing the notation $\psi_3 = \psi_1 \boxplus \psi_2$. The calculations based on the exact solutions for the three potentials must and will show that the statistical averaging procedure fails. We could in this respect make detailed comparisons between the calculations. This leads to the astounding conclusion that the averaging procedure is flawed.

This is certainly counter-intuitive, but we must point out that we are not used to logic that allows for undecidability. The averaging procedure is flawed in ternary logic. We can render this intelligible by pointing out that there does not exist a *common* probability distribution that could be used for the three experiments (with potentials V_1, V_2 and V_3). In fact, in the double-slit experiment with the potential V_3 there exist probabilities for an answer "do not know" to the question: "did the particle travel through slit S_j?". Such probabilities do not exist in the two other experiments, where only one slit is open and the answers can only be "yes" or "no". Due to the absence of a common probability distribution for the three experiments, one cannot calculate the diffraction pattern of the double-slit experiment by averaging over the probability distributions of the single-slit experiments. That the averaging procedure is flawed in ternary logic is confirmed by the empirical evidence from physics and by what quantum mechanics itself tells about probability distributions (as will be discussed in the very important Footnote 31). This will be further elaborated in Section 12 and illustrated in Fig. 3.

Whereas Kleene, Priest and Łukasiewicz have proposed truth tables for ternary logic (see reference [18]), it is not clear to us if anyone has ever derived rules for calculating probabilities according to ternary logic from first principles. We therefore do not know how the probabilities must be calculated within such logic in order to compare them with the quantum mechanical results. Quantum mechanics is all we have to calculate such probabilities in a very specific context (that relies e.g. on the existence of spin). This way, the problem becomes even more fundamental, because it evolves from a probability paradox to a paradox of ternary logic.

The paradox results from the fact that we try to understand ternary logic from a viewpoint based on binary logic. One could compare this to trying to understand hyperbolic geometry from the viewpoint of Euclidean geometry. Just as hyperbolic and Euclidean geometry are incompatible axiomatic systems, ternary and binary logic are incompatible axiomatic systems. There is thus absolutely no ground for using binary intuition to tackle problems with ternary logic, but of course we could try our luck. For example, we might guess that we can obtain probabilities in ternary logic by describing the ternary statistical ensemble as a constructed statistical ensemble of all possible binary ensembles we can obtain by redistributing the undecided outcomes over the two decided outcomes and averaging over this constructed ensemble, like we do in binary logic when we do not have all the information. To show that this intuition is wrong, nothing is better than giving a counterexample. The counterexample is the double-slit experiment where clearly the probability is not given by $|\psi_1|^2 + |\psi_2|^2$ but by $|\psi_3|^2 \approx |\psi_1 \boxplus \psi_2|^2$, where the index 3 really refers to the third option. It is then useless to insist any further.

We may feel that binary logic is different from ternary logic in that it is so obvious that it would not need any experimental evidence from physics to justify its rules. But the reason why we find it obvious might be that we have only been confronted in our human evolution with macroscopic physical evidence where binary logic prevails. Moreover, the idea of undecidability seems to connect neatly to Heisenberg's idea of uncertainty. Eventually, this paradox is not any more harsh than the paradox of Banach and Tarski in measure theory [19].

There is another way to present the problem of a contextual situation. We can think of checking that the history associated with ψ_1 and takes the particle through slit S_1 is truly decided by taking a path ψ^* backwards in time. If the context is truly decided, all backward paths must die on slit S_2, such that $\psi_1\psi^* = \psi_1\psi_1^*$. If the context is truly undecided, the backward paths can thread through both S_1 and S_2. Moreover, the phases must fulfill a compatibility condition. We have then $\psi_1\psi^* = \psi_1\psi_1^* + \psi_1\psi_2^*$.

You can however, not consider this term in an isolated way in its own right, because the context is undecided. You must include the analogous reasoning for $\psi_2\psi^*$. The compatibility issue leads exactly to the wave functions we proposed in reference [1]. It also contacts truly nicely with Cramer's handshake mechanism in his transactional interpretation of quantum mechanics [20]. The waves are traveling backwards in time but they are only a mathematical expedient to check the self-consistence of the wave function when it is non-local. They do not correspond to physical reality. They are just a way to deal with the non-locality of the wave function and the set-up. There is no true violation of causality in the wave function. Probabilities can only be approved as good probabilities for the non-local problem if all the handshakes have been carried out.

8 Justification of the quantum rules for the calculation of the probabilities of coherent and incoherent scattering

Whereas these remarks (and especially those in Footnote 31 and in Fig. 3) solve the probability paradox conceptually, they do of course not explain the textbook rule that tells us how we must calculate probabilities in the coherent case. But that the paradox can be solved conceptually is already an important result. It must be clear from the analysis given above that we can have actually two types of sums $\psi_1^{(coh)} \boxplus \psi_2^{(coh)}$ and $\psi_1^{(incoh)} + \psi_2^{(incoh)}$. As explained, we can introduce the incoherent sum rule by mere definition and it corresponds to the prescriptions of quantum mechanics when we apply the superposition principle. The coherent rule cannot be justified this way. It has to be done differently. We must eventually justify it by a different principle.

In the case (Q2), the solution of the Dirac equation may be a state $\psi = \psi_1 \boxplus \psi_2$. In such a state ψ, the probability is a quadratic expression $p = |\psi|^2$ in terms of the spinor ψ. This is because probabilities are components of a four-vector, viz. the probability charge-current four-vector. In fact, for every quantum mechanical equation in a textbook, the introduction of the wave equation is followed by a derivation of a continuity equation for the probability charge-current four-vector from it. It is this derivation that justifies using $p = |\psi|^2$ or $p = \psi^\dagger\psi$ as a probability. The expression for the probability charge-current four-vector in the Dirac equation shows that it has exactly the symmetry of a four-vector. Group theory shows that four-vectors must be covariant bilinear expressions in terms of the spinors. We can consider the case of the Schrödinger equation as derived from the Dirac equation. As we have mentioned, the wave functions that occur in the Pauli and Dirac equations are sums of spinors, because the eigenfunctions ψ of a spin operator $\hat{S}$ are always sums of spinors. The reason for this is that the spin operator is a reflection operator and reflections do not have eigenfunctions on the group. They only have eigenfunctions on the group ring. This special sum rule is not at all a result of the group theory, because the group theory does not even provide a meaning for such linear combinations. As far as the group theory is concerned, reflection operators have no eigenfunctions on the group and the result we obtain by carrying out the algebra mindlessly is meaningless, because in doing so we fail to acknowledge for the fact that the group is not a vector space but a manifold. But for these sums that represent spin operator eigenfunctions we can justify the rule for coherent summing because they lead to the Pauli and Dirac equations, and for these equations the continuity equations derived from them show that that $\psi^*\psi$ and $\psi^\dagger\psi$ behave as probabilities. But this special case cannot be used to justify the sum rule in the double-slit experiment. It can only be justified by proving that $\psi = \psi_1 \boxplus \psi_2$ is to a good approximation a solution for the wave equation.

The quadratic link between four-vectors and spinors can only be derived from the group theory for pure spinors describing group elements, not for linear combinations of pure spinors. However the fact that we can derive a continuity equation from the wave equation shows then that such a quadratic link also exists for some special linear combinations of spinors, viz. those that are eigenvectors of the spin operator and the solution $\psi = \psi_1 \boxplus \psi_2$ of the wave equation for the double-slit experiment.

This proposal is extremely subtle. It is indeed so subtle that it may look like farfetched faultfinding to many people, but it must be realized that the experimental results just leave us with the impression that we are unable to think straight. What the hell would be wrong with assuming that the electron travels either through slit S_1 or slit S_2. According to our proposal, this assumption can be correct, but the statistical method fails to reproduce $|\psi|^2 = |\psi_1 \boxplus \psi_2|^2$ as can be checked from the calculations, such that this method must be wrong. The subtle arguments show that there is indeed no reason to assume that the statistical method would work because ψ_1 and ψ_2 are wave functions obtained from different potentials (V_1 and V_2) than V_3. Therefore they are not the appropriate functions to split up ψ_3 into $\psi_3(S_1) + \psi_3(S_2)$. With this proposal we can at least reconquer the impression that we are able to think straight.

9 Justifying the coherent sum rule $\psi = \psi_1 \boxplus \psi_2$

9.1 By using the Fourier transform and/or the Born approximation

We want now to discuss why $\psi = \psi_1 \boxplus \psi_2$ will often be a good approximation of the wave function. Here we will discuss how we can invoke the Born approximation and/or the Fourier transform to justify this idea. In the next section we will refine this by using Huyghens' principle from optics. In Subsection 9.3, we will improve this further by specifying the Huyghens' principle for quantum mechanics. The general form of a wave in the experiment is $\psi(\mathbf{r}) = a(\mathbf{r})e^{\imath\mathbf{k}\cdot\mathbf{r}}$, whereby we note $(x, y) = \mathbf{r}$, $(k_x, k_y) = \mathbf{k}$. The reason why we write $a(\mathbf{r})$ instead of just 1 is as follows. We will consider plane waves. This is an approximation as the source is in reality a small surface. But when the source is very far away from the slit, it can be treated as though it were a point source. The waves are then spherical. However the spherical wave front that arrives at the slit will locally, i.e. on the length scale of the slit, just be indistinguishable from the tangent plane to the spherical wave front. E.g. for plane waves propagating along the y-direction we would thus have $k_x = 0$, $k_y = |\mathbf{k}| = k$. Taking the idea of plane wave literally, we must consider the source as a line $y = -d_G$. We will first assume that the phase of the wave is zero when the particle is emitted by the source. A particle that arrives in a point $(x, 0)$ of the plane of the set-up must then have left the source in $(x, -d_G)$. The amplitude of a plane wave in the Oxy-plane is just $a(\mathbf{r}) = 1, \forall \mathbf{r} \in \mathbb{R}^2$. The wave is then $\psi(\mathbf{r}) = e^{\imath k(y+d_G)}$. The values in the plane of the slit are thus $\psi(x, 0) = e^{\imath d_G k}$. However, the potential imposes boundary conditions on the wave function. The wave function must vanish in all points where $V(\mathbf{r}) = \infty$. This does not need to be true when the particles are reflected by the device, but we are interested only in the particles that are transmitted through the device. The points $(x, 0) \parallel V(x, 0) = \infty$ do not transmit. In all other points the wave just "sees" free space, such that the local solution of the wave equation in these points will thus just be the initial plane wave. The boundary conditions are not compatible with the plane-wave *Ansatz* because the *Ansatz* just implies $a(\mathbf{r}) = 1$. Combined with the assumption that a particle that arrives at $(x, 0)$ is emitted at $(x, -d_G)$ this forces us to assume that not all points of the source on the line $y = -d_G$ are emitting. The incoming wave is then not $\psi(\mathbf{r}) = e^{\imath k(y+d_G)}$, but $\psi(\mathbf{r}) = C(x)\, e^{\imath k(y+d_G)}$, where:

$$\begin{aligned} C(x) &= 1, \quad \forall x \in [-D/2 - w/2, -D/2 + w/2] \cup [D/2 - w/2, D/2 + w/2], \\ C(x) &= 0, \quad \forall x \in \mathbb{R}\backslash([-D/2 - w/2, -D/2 + w/2] \cup [D/2 - w/2, D/2 + w/2]). \end{aligned} \tag{16}$$

We are thus modulating the wave function with $C(x)$. We have then $\psi(x, 0) = e^{\imath d_G k} C(x)$. In all these equations $k = \frac{2\pi}{\lambda}$ contains the information about the energy of the incoming particle. To get rid of the term $e^{\imath d_G k}$, we will assume from now on that the particles leave the source with the phase $e^{-\imath d_G k}$, such that for $y > 0$ the phase of the wave function is just $e^{\imath k_y y}$. We will further modify this. We will consider that the incoming wave has not the form $e^{\imath k_y y}$ but:

$$\psi(x, y) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{k_x^2}{2\sigma^2}}\, e^{\imath(k_x x + k_y y)}. \tag{17}$$

This implies that the distribution of the momentum $p_x = \hbar k_x$ is not a Dirac measure $\delta(p_x)$ positioned at $p_x = 0$, but a Gaussian distribution:

$$G(k_x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{k_x^2}{2\sigma^2}}, \tag{18}$$

centered around this value with a width $\hbar\sigma$, which is certainly more realistic from a physical viewpoint.[26] The transmission amplitude $T(k_x)$ will become:

$$T(k_x) = \int_{-\infty}^{\infty} C(x) \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{k_x^2}{2\sigma^2}}\, e^{\imath k_x x}\, dx. \tag{19}$$

This expresses that only in the points where the potential is zero the particle is transmitted. The factor $\frac{1}{\sigma}\, e^{-\frac{k_x^2}{2\sigma^2}}$ can be put in front of the integral. The Fourier transform of $C(x)$ is given by:

$$\mathscr{F}(C)(k_x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} C(x)\, e^{\imath(k_x x)}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-D/2-w/2}^{-D/2+w/2} e^{\imath k_x x}\, dx + \frac{1}{\sqrt{2\pi}} \int_{D/2-w/2}^{D/2+w/2} e^{\imath k_x x}\, dx. \tag{20}$$

Here each term:

$$\frac{1}{\sqrt{2\pi}} \int_{\varsigma D/2-w/2}^{\varsigma D/2+w/2} e^{\imath k_x x}\, dx = \frac{\imath w}{\sqrt{2\pi}} J_0\!\left(\frac{k_x w}{2}\right) e^{\imath \varsigma k_x D/2}, \tag{21}$$

with $\varsigma \in \{-1; 1\}$, represents the contribution ψ_j of a single slit S_j. When both slits S_1 and S_2 are open, the term:

$$e^{\imath k_x D/2} + e^{-\imath k_x D/2} = 2\cos\frac{k_x D}{2}, \tag{22}$$

will occur as a pre-factor of the terms $\imath \frac{w}{\sqrt{2\pi}} J_0(\frac{k_x w}{2})$ and represent the interference. With the pre-factor $\frac{1}{\sigma}\, e^{-\frac{k_x^2}{2\sigma^2}}$ we recover completely the expression of the initial width distribution $G(k_x)$ at the end of the calculation. The term $\imath w J_0(\frac{k_x w}{2})$ will fulfill the rôle

[26] We could qualify this as a "wave packet" but one should not identify the electron with a physically meaningful wave packet. The electron is and remains a point particle. We conceive the wave packet rather as a heuristic mathematical tool. It is a more appropriate tentative solution of the wave equation than the Dirac measure in view of the global character of the boundary conditions.

of a hull function for the interference term. The final amplitude is thus $\imath\, 2w \cos \frac{k_xD}{2} J_0(\frac{k_xw}{2})\, G(k_x)$. This calculation yields thus exactly the rule $\psi = \psi_1 \boxplus \psi_2$.

The final amplitude $\mathscr{F}(C)$ is related to the Fourier transform $\mathscr{F}(V)$ of the potential in a 1-1 fashion. The amplitude is not exactly the Fourier transform of the potential V which takes values ∞ in some points $\mathbf{r}$ and 0 in some points $\mathbf{r}'$. It is the Fourier transform of a related function C which takes values 0 in the very same points $\mathbf{r}$ and 1 in the very same points $\mathbf{r}'$. We see thus that the probability amplitudes $C(k_x)$ are the Fourier transform of the "potential" C. This is at least in its spirit in agreement with the Born approximation:

$$\frac{d\sigma}{d\Omega} = \frac{m_0}{4\pi^2}\, |\mathscr{F}(V_s)(\mathbf{q})|^2. \tag{23}$$

for the differential cross section that describes the scattering of a particle with mass m_0 by a scattering potential V_s, where $\mathbf{q}$ is the momentum transferred in the scattering. The result we obtained can be related to the result obtained in the Born approximation. The relation between V and C is that we must renormalize the infinite values that V takes to 1. This yields V_1. And then we must take $C = 1 - V_1$. It is more rigorous to formulate this the other way around by considering that the idealized double-slit potential V is given by:

$$\forall x \in \mathbb{R}: \;\; V(x) = \lim_{V_0 \to \infty} V_0 V_1(x). \tag{24}$$

This can be related to a Born-type scattering potential V_s as follows:

$$\forall x \in \mathbb{R}: \;\; V_s(x) = V_0(1 - V_1(x)) = V_0 C(x), \quad C(x) = \lim_{V_0 \to \infty} \frac{V_s}{V_0} = 1 - V_1(x). \tag{25}$$

We see then that Born describes the scattering by a potential V_s that has the form of two identical mountains in an empty landscape. The relation is illustrated between C and V_s is illustrated in figure 2. The Fourier transform of 1 in $1 - V_1(x)$ leads just to a Dirac measure, that we did not reproduce in our geometrical calculation of Eqs. 17-22. In Born's problem, a lot of electrons will not be scattered at all and yield a Dirac measure $\delta(p_x)$ that accounts for the difference between the result we found for C from the geometrical calculation in Eqs. 17- 22 and Born's result for $V_s = V_0 C$. The scattering problem for the potential V_s (when $V_0 \to \infty$) is the "negative" of the transmission measurement for the potential V. The scattering problem is physical, while the transmission experiment is geometrical. We can use one to calculate the other because the total transmission of the sum of the potentials V_s and V will be zero when $V_0 \to \infty$.

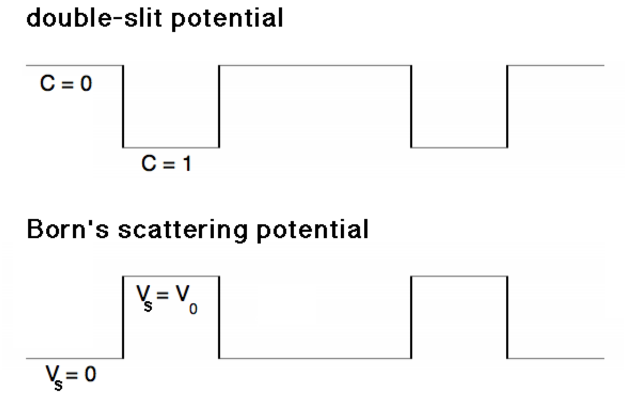

Fig. 2. Comparison of the double-slit potential V (with its values $V(x) \in \{0, \infty\}$ labeled in terms of $C(x) \in \{0, 1\}$) and Born's scattering potential V_s that corresponds to it.

We can thus solve the Schrödinger equation for a double-slit experiment exactly without any use of the superposition principle, and calculate the scattering. In principle the Born approximation will yield a good approximation to this scattering. It is this Born approximation that will obey the pseudo "superposition principle" that we just must sum the two single-slit wave functions of the contributions of the two slits. The principle is not a real superposition principle but a Huyghens' principle. For the more rigorous true solution, we cannot prove this sum rule. In fact, we have in our approach presented the experiment as a geometrical transmission experiment rather than as a physical scattering experiment, and without introducing the Gaussian width distribution, we could never have drawn in the Fourier transform into our calculation. The exact identity between the result of the geometrical calculation for C and the result of the physical calculation for the related potential V_s in the Born approximation can only underline that the result of the geometrical calculation is also an approximation (see also Footnote 24). This is in our opinion a way to justify that the sum rule yields a good, but not exact description of the wave function in a double-slit experiment.

The calculation also confirms the thesis announced in Section 6 that we are measuring the set-up with electrons, rather than measuring electrons with the set-up, because the wave function is the Fourier transform $\mathscr{F}(V)(\mathbf{q})$ of the potential $V(\mathbf{r})$. The information contained in $\mathscr{F}(V)(\mathbf{q})$ is the same as the information contained in $V(\mathbf{r})$. This is especially obvious because the Fourier transform of $\mathscr{F}(V)(\mathbf{q})$ is again $V(\mathbf{r})$. The amplitude of the wave function contains thus nothing more and nothing less than the complete information about the experimental set-up. The electrons are providing so to say a photograph, not of the set-up, but of the Fourier transform of the set-up, which is an equivalent representation of the information. It is the thesis from Section 6 that renders the double-slit experiment intelligible. It provides a strong motivation for Bohr's interpretation that we must take into account the rôle of the measuring device in analyzing physical experiments. The result also provides evidence for the thesis that the Dirac equation is classical and that it is the way we solve it that renders it quantum mechanical.

Finally, it also can be used to clarify something that is really mysterious, viz. that to solve certain problems we must postulate that the wave function must be a (single-valued) function.[27] This postulate leads to a quantization condition in a straightforward manner. This is hard to justify if we think about quantum mechanics as a set-up measuring electrons. It is much more easy to accept if we think about quantum mechanics as electrons measuring a set-up, because for this set-up we must then find a "probability amplitude function" (in the form of a Fourier transform) that describes it unambiguously and its is only reasonable that this should be a function, because the Fourier transform of a function is a function. The wave function must be global, and be assembled from its definition over local patches of its definition domain. We knit the patches together and in order to avoid any possible internal contradictions in the global assembly, we must make sure that calculations based on alternative paths through the definition domain lead to consistent results. The histories must be consistent. Therefore we use the Huyghens' principle and require the wave function to be a single-valued function. This way we can make sure that (e.g. in a Dirichlet problem) the solution proposed to satisfy the local boundary conditions in one place does not clash on a global level with the solution proposed to satisfy the local boundary conditions in another remote place. From this point of view, it is not at all obvious that the existence of a global solution would be granted, and there is then no surprise that this can lead to quantization.[28] The thesis advocated in Section 6, is thus a kind of paradigm shift as described by Kuhn [21], whereby all at once the drawing of a duck changes into the drawing of a rabbit.

The calculation also illustrates the idea how we one can generalize the Dirac equation from a path $\mathcal{P}$ to $\mathbb{R}^4$ in the presence of a potential, by just generalizing it first to $\mathbb{R}^4$ and then imposing the potential. However, the calculation we described here is based on the introduction of a packet of waves with central symmetry. On the local scale of the slits, we have replaced the plane wave which is the exact generalization from $\mathcal{P}$ to $\mathbb{R}^4$, by a radially propagating wave. This is thus a kind of cheat. (We have further modulated this radially propagating wave with a Gaussian, but fortunately this modulation does not intervene in the calculation above). Without this cheat, we would never have been able to make the Fourier analysis, because the term $k_x x$ would have been absent in the exponential. Furthermore, with a narrow incoming beam we will not reproduce the full span of the diffraction pattern behind the slits, which seems to bend around the corner. This is because we describe the experiment by pure physical transmission with a highly idealized potential. We can try to fiddle with the width of the Gaussian to obtain a reasonable result. But we can appreciate that this is all sketchy and intuitive rather than hard proof. We should actually not expect to be able to render the argument any more rigorous because the tentative expression $\psi_1 \boxplus \psi_2$ is not rigorous. Taking it literally would lead to insuperable conceptual difficulties (see Footnote 32). What is exact is the extrapolation of the wave equation from $\mathcal{P}$ to $\mathbb{R}^4$ leading to the boundary conditions expressed in Eq. 16, not the approximative solution for that extrapolated equation we tried to find here in terms of presumably real histories. We loose our grip on the detailed real histories in the approximation $\psi_1 \boxplus \psi_2$. We will see that the wave function can allow for a broad fan of incoming angles, if we do not consider these angles as physical and only occurring within a purely mathematical Huyghen's principle. First we will describe Kirchoff's Huyghens' principle, then we will describe Feynman's improved Huyghens' principle in the form of his path integral method. With Kirchoff's approach one cannot justify the generalization of the definition domain of the wave equation. It takes the generalized wave equation as its starting point, and the wave equation in question applies to photons rather than to electrons. It illustrates how the same general principles can show up in different situations, and lead to the same conclusions, while certain details can be actually different. But Feynman's approach can be used to justify the generalization of the definition domain as Feynman has shown that his Huyghens' principle leads to the Schrödinger equation. In both approaches we will see that it is crucial to have the electrons explore the full set-up, even if this must be considered as a purely mathematical description that does not need to correspond to the real physical situation. What counts in the physical situation is that we can state that the wave function contains the complete information

[27] This intervenes in the solution of the wave equation for the spectrum of the hydrogen atom, as discussed in [1], p. 206, p. 281. It also intervenes in the discussion we made in reference [1] of the double-slit experiment (see pp. 327-335).

[28] We could object that we are only interested in the probabilities such that one could allow for inconsistencies in the phase. But the phase is also important because the spinning and orbital motion are coupled in quantum mechanics. When a sub-atomic particle rotates faster, then its rest mass will be increased. This in turn will change its orbit. And due to Thomas precession effects, changing the orbit will change the rate of the spinning motion and the rest mass. Postulating that the wave function must be a function permits to treat this coupling by defining unambiguously the frequency of the spinning motion through the knowledge of the exact value of the phase. Electrons do have a phase, such that the phase is a quantity with a physical meaning. It is the rotation angle (modulo 4π) of the spinning electron. The phase is what makes a spinor different from a vector. Furthermore, if we dropped the phase from the wave function, the result $|\psi|$ would no longer constitute the full information about the set-up. That is, from $|\mathscr{F}\,[V(\mathbf{k})]\,|$ we can no longer recover $V(\mathbf{r})$ by the inverse Fourier transform, like we can from $\mathscr{F}[V(\mathbf{k})]$. This e.g. well-known in crystallography where it gives rise to a phase problem and the introduction of so-called Patterson functions.

about the set-up, which is expressed by (the non-physical) mathematical image that the electrons visit all parts of the set-up even if this takes very counter-intuitive paths. These are the elaborations described in the following sections.

9.2 Photons in optics: By using the historical Huyghens' principle

If a wave equation allows for a Huyghens' principle we can argue that the solution $\psi = \psi_1 \boxplus \psi_2$ should be a very good approximation. For light waves we can e.g. use a principle due to St. Venant discussed in reference [22] (pp. 203-204). The idea is to put the wave function and its derivatives zero on the diffracting screen (exactly as we argued for the infinite potential). As Longhurst is pointing out, this leads to problems at the points that separate the regions where $V = \infty$ and $V = 0$ as in these points the continuity conditions are not satisfied. Let us however assume that we can neglect this problem. By using Venant's method we can derive then $\psi = \psi_1 \boxplus \psi_2$.

However, this derivation is not fully exact for another reason. In St. Venant's method the wave propagation must be strictly forward. But this is not true in the Huyghens principle for light rays. The derivation of Huyghens' principle for light rays by Kirchoff is also discussed in reference [22]. In this Huyhens' principle there is an obliquity factor that reduces the weight of backwards traveling waves. Nevertheless, it does not completely rule out backwards traveling waves. St. Venant's principle will however only prove the rule $\psi = \psi_1 \boxplus \psi_2$ exactly if we exclude the possibility of backwards traveling waves all together. Else, we can consider for the potential V_3 (when the two slits are open) a path C_1PC_2Q where P is situated before the slits (such that on C_1P the wave is traveling backwards) and Q is situated behind the slits (such that on PC_2Q the wave is traveling forwards). Such a path does not exist for the potentials V_1 and V_2, which clearly indicates that the solution $\psi = \psi_1 \boxplus \psi_2$ is not exact. Here the waves should not be considered as physically meaningful. They are just a mathematical tool that occurs in the Huyghens' principle for the wave equation. The Huyghen's principle itself is also just a mathematical tool because it contains features that cannot be justified physically, such as the obliquity factor, a weighting factor $\frac{1}{\lambda}$ and a phase difference of $\frac{\pi}{2}$ between primary waves and secondary waves. There is no dictionary that would enable to translate Kirchoff's mathematics in a 1-1 fashion to meaningful physics for light rays. But we may present some wrong physical interpretation for it. This will exhibit counterintuitive aspects because it is wrong, but it could be helpful to present pseudo-physical pictures for the calculations that have great mnemonic value. This could be true for other Huygens' principles as well. Feynman's path integral method can also be considered as a Huyghens' principle and must also be considered as a purely mathematical tool. It is in this respect that we can understand what Feynman reported e.g.: *"...there is also an amplitude for light to go faster (or slower) than the conventional speed of light. You found out in the last lecture that light doesn't only go in straight lines; now, you find out that it doesn't only go at the speed of light!"* [23], which is completely incomprehensible if taken literally.

9.3 The path integral method as the exact Huyghen's principle for quantum mechanics

We derive here a formula given by Dirac [24], that has been used by Feynman as the starting point in his development of the path integral method [25]. This equation is a form of the Huyghens' principle. In the rest frame of the electron $\psi(\boldsymbol{\rho}, \tau) = e^{-\frac{\imath}{\hbar} m_0 c^2 \tau}$, such that:

$$\psi(\boldsymbol{\rho}, \tau') = \psi(\boldsymbol{\rho}, \tau)\, e^{-\frac{\imath}{\hbar} m_0 c^2 (\tau' - \tau)} \tag{26}$$

and

$$\psi(\boldsymbol{\rho}', \tau') = \int_{\mathcal{B}} \psi(\boldsymbol{\rho}, \tau)\, e^{-\frac{\imath}{\hbar} m_0 c^2 (\tau' - \tau)}\, \delta(\boldsymbol{\rho}' - \boldsymbol{\rho})\, d\boldsymbol{\rho} \tag{27}$$

for any set $\mathcal{B}$ that contains $\boldsymbol{\rho}'$. Here $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ are positions in the rest frame of the electron. With $\mathcal{B} = \mathbb{R}^3$ we can take the following form for the Dirac measure:

$$\delta(u) =_D \lim_{\alpha \downarrow 0} \frac{1}{2\sqrt{\pi \alpha}} e^{-u^2/4\alpha} \tag{28}$$

where the subscript D will serve to remind us that we are rather dealing with an equivalence in the sense of distributions. By putting:

$$u = \xi' - \xi; \qquad \frac{\imath\, m_0}{2\hbar(\tau' - \tau)} = -\frac{1}{4\alpha} \tag{29}$$

where $\boldsymbol{\rho} = (\xi, \eta, \zeta)$, we will obtain:

$$\delta(\xi' - \xi) =_D \lim_{\tau' - \tau \to 0} \sqrt{\frac{m_0}{2\pi\hbar\imath(\tau' - \tau)}} \times \exp\left[\frac{\imath\, m_0}{2\hbar} \left(\frac{\xi' - \xi}{\tau' - \tau} \right)^2 (\tau' - \tau) \right] \tag{30}$$

and in three dimensions:

$$\delta(\boldsymbol{\rho}' - \boldsymbol{\rho}) =_D \lim_{\tau'-\tau\to 0} \left(\frac{m_0}{2\pi\hbar\imath(\tau'-\tau)} \right)^{\frac{3}{2}} \exp\left[\frac{\imath\, m_0}{2\hbar} \left(\frac{\boldsymbol{\rho}'-\boldsymbol{\rho}}{\tau'-\tau} \right)^2 (\tau'-\tau) \right] \tag{31}$$

If we work non-relativistically, for a free particle the Lagrangian is given by $\mathscr{L}(\mathbf{r},\mathbf{v}) = \frac{1}{2}\, m_0 v^2$. Hence $\frac{1}{2} m_0\, [\, \frac{\rho'-\rho}{\tau'-\tau}\,]^2 \cdot (\tau'-\tau)$ can be written as $\mathscr{L}\,[\,\boldsymbol{\rho}, \frac{\rho'-\rho}{\tau'-\tau}\,] \cdot (\tau'-\tau)$, such that

$$\delta(\boldsymbol{\rho}' - \boldsymbol{\rho})\, e^{-\frac{\imath}{\hbar} m_0 c^2 (\tau'-\tau)} =_D \lim_{\tau'-\tau\to 0} \left(\frac{m_0}{2\pi\hbar\imath(\tau'-\tau)} \right)^{\frac{3}{2}} \exp\left[\frac{\imath}{\hbar}\, \mathscr{L}\,(\boldsymbol{\rho}, \frac{\boldsymbol{\rho}'-\boldsymbol{\rho}}{\tau'-\tau}\,) \cdot (\tau'-\tau) \right] \tag{32}$$

which after substitution into Equation (27) yields:

$$\psi(\boldsymbol{\rho}', \tau') = \lim_{\tau'-\tau\to 0} \int_{\mathbb{R}^3} \psi(\boldsymbol{\rho}, \tau) \left(\frac{m_0}{2\pi\hbar\,\imath\,(\tau'-\tau)} \right)^{\frac{3}{2}} \exp\left[\frac{\imath}{\hbar}\, \mathscr{L}\,(\boldsymbol{\rho}, \frac{\boldsymbol{\rho}'-\boldsymbol{\rho}}{\tau'-\tau}\,) \cdot (\tau'-\tau) \right] d\boldsymbol{\rho} \tag{33}$$

Introducing an instantaneous Lorentz transformation that maps: $\tau' - \tau \mapsto \varepsilon = t' - t$, $(\tau, \tau') \mapsto (t, t')$, $(\boldsymbol{\rho}, \boldsymbol{\rho}') \mapsto (\mathbf{r}, \mathbf{r}')$, we obtain Dirac's formula:

$$\lim_{\varepsilon\to 0} \psi(\mathbf{r}', t+\varepsilon) = \lim_{\varepsilon\to 0} \left(\sqrt{\frac{m_0}{2\pi\hbar\,\imath\,\varepsilon}} \right)^3 \int_{\mathbb{R}^3} \psi(\mathbf{r}, t)\, \exp\left[\frac{\imath}{\hbar}\, \mathscr{L}(\mathbf{r}, \frac{\mathbf{r}'-\mathbf{r}}{\varepsilon})\, \varepsilon \right] d\mathbf{r} \tag{34}$$

where we have used the Lorentz covariance of $\mathscr{L}$ which continues to express the proper time, such that we are allowed to extrapolate the formula to the motion of a particle in a potential. Introducing the potential this way changes the nature of the extrapolation from geometrical to physical. In fact, the Eq. 29 we started from is purely geometrical. Eq. 34 is in a sense an integral form of the equation $\frac{d}{d\tau}\psi = \frac{\imath m_0 c^2}{\hbar}\psi$. The fact that the fully relativistic wave function should have four components as exemplified by the Dirac equation, shows why Eq. (34) can only have a limited domain of validity (In fact, there is no exact instantaneous Lorentz transformation of the type we postulated). Note also that our expression contains the right normalization constant and that in the classical Lagrangian the rest energy is ignored. Feynman had to introduce the normalization constant *a posteriori* into Dirac's formula in order to obtain the correct expressions. Feynman has shown that Equation (34) leads directly to the Schrödinger equation [25]. The Schrödinger equation can also be derived from the Dirac equation. We should also not too much surprised that in Feynman's method a particle seems to take all possible paths. As the information we obtain about the particle is rather time-like, there is very little information about the path the particle has taken: In free space, in the rest frame of the particle, there is none! In a hand-waving fashion we can argue that a way to reproduce total absence of information about the paths is to assume that all paths occur with the same weight. More rigorously it must be a consequence of a finding by Hadamard that for certain partial differential equations the solutions exhibit a Huyghens' principle [26].

9.4 Summary

The three approaches in Subsections 9.1-9.3 are not rigorous in all their details, but they all converge to the same general idea that one can generalize the Dirac equation in the presence of a double-slit potential from a single path $\mathcal{P}$ to $\mathbb{R}^4$. The wave function over this generalized definition domain is obtained by considering other alternative paths through this extended domain, which are consistent histories. These alternative paths cover the entire domain. The solution of the wave equation over this extended domain is a wave function which contains the complete information about the experimental set-up (e.g. in the form of the Fourier transform of the potential) in a one-to-one correspondence. It is this simultaneous extension of the equation and its solution by a Huyghens' principle that permits to understand the coherent sum rule and the so-called "particle-wave duality". The latter does not imply that the electron would sometimes behave like a particle and sometimes like a wave. An electron always behaves like a particle. What behaves like a wave is the wave function, i.e. the full information about the set-up collected by a statistical ensemble of electrons used to study the set-up.[29] That the information collected about the set-up is not presented directly but

[29] It may be noted in this respect that even in a classical water wave the individual water molecules do not move in wavelike manner. They only display some local circular motion (see e.g. Figure 6 from reference [27]). The wave behaviour is also here a property of an ensemble of a large number of molecules rather than of the individual molecules. The analogy stops here because the water molecules are in mutual interaction, while in a carefully designed double-slit experiment there may be only one particle present at the time. Moreover, electrons do move over large distances. They do have a phase due to their spinning motion and mathematically this does look like a wave on $\mathcal{P}$. However, the spinor ψ that describes the spinning motion is a temporal wave, not a wave that truly propagates in space (as indicated by its phase velocity $c^2/v > c$, see also reference [1], pp. 199-205). The extrapolation of ψ from $\mathcal{P}$ to $\mathbb{R}^4$ carries the local wavelike behaviour on $\mathcal{P}$ over to a global wavelike behaviour on $\mathbb{R}^4$. It is the fact that the wavelike appearance just expresses the spinning motion which explains that the probability waves can propagate in vacuum. The propagation in vacuum was historically considered as a riddle for electromagnetic waves.

e.g. under the form of a Fourier transform of the potential, is due to the fact that an electron has a phase associated with its spinning motion. It is this spinning motion that makes the representation of the information collected about the set-up look like a wave. It turns out that this wave is a probability amplitude (see Section 10). The procedure followed here to generalize the definition domain is case-specific. The generalizations must be discussed on a case-by-case basis. E.g. the generalization of the wave function for the energy spectrum of the hydrogen problem must be made in a completely different way. It is also lengthy and involved. It is based on the result announced in Footnote 7 and on Kepler's area theorem (see reference [1]). Tunneling requires yet another approach.

10 A note on the probability densities

The derivations of the expressions for the four-vector $j_\mu \equiv (c\rho, \mathbf{j})$ that gives the probability charge-current density from the free-space Dirac equation or the Schrödinger equation are well known (see e.g. reference [28]). Using the same methodology for the Klein-Gordon equation we find a continuity equation for the charge-current-like four-vector:

$$j_\mu = \imath\,(\psi^*\,\partial_\mu\psi - \psi\,\partial_\mu\psi^*), \tag{35}$$

such that:

$$c\rho = \imath(\psi^*\,\frac{\partial\psi}{\partial ct} - \psi\,\frac{\partial\psi^*}{\partial ct}). \tag{36}$$

As $\psi^*\,\frac{\partial\psi}{\partial ct} - \psi\,\frac{\partial\psi^*}{\partial ct} = [\,\psi^*\,\frac{\partial\psi}{\partial ct}\,] - [\,\psi^*\,\frac{\partial\psi}{\partial ct}\,]^*$, $c\rho \in \mathbb{R}$ is real. However, $c\rho$ can be negative, such that it cannot be defined as a probability density. The reason for this failure is the fact that the Klein-Gordon equation is of second order. To know its solution we must know both ψ and $\frac{\partial\psi}{\partial t}$ at some moment in time. The fact that $\frac{\partial\psi}{\partial t}$ helps in defining the solution makes then its way into the Eq. 35.

The equation for electromagnetic waves propagating in free space is a special case of the Klein-Gordon equation for $m_0 = 0$. One might consider to resolve the problem of deriving an expression for the probability density for electromagnetic waves by taking the square root of the equation for electromagnetic waves in free space. This would lead to a Dirac-like equation whereby the rest mass m_0 of the particle is zero. From this equation we could then derive an expression for the probability density just like for the Dirac equation. But this runs contrary to the idea that photons must have scalar wave functions. It follows that the rule that $p = |\psi|^2$ must be derived in a completely different way for light waves in free space. The derivation can be found in the Feynman Lectures [29] (paragraphs 27.2 -27.4). As the energy of the electromagnetic field is a measure for the number of photons (which for our purposes we assume here to be all of the same energy), we can consider this energy to be proportional to the probability density. The Poynting vector is then a measure for the energy flow.

We mentioned that the equation for electromagnetic waves propagating in free space is a special case of the Klein-Gordon equation whereby $m_0 = 0$. The continuity equation Eq. 35 is therefore an entirely correct result for light waves. However, we have no physical meaning for it. The waves are only mathematical tools to describe the probability distribution defined by a specific set-up. (In the Dirac equation, the wave has originally physical meaning because it describes the spinning motion of the electron over a physical path, but this meaning does not apply to the extrapolation of the wave function from this physical path to a definition domain that corresponds to the whole of $\mathbb{R}^4$. It is this extrapolated function we use when we solve the equation). The quantity j_μ corresponds thus to some mathematical flow that occurs in the wave function ψ but has no physical meaning in terms of physical quantities like energy or mass. It resembles in this sense the mathematical flow that occurs in Huygens' principle which also has no true physical meaning. We may note that the term $\nabla\cdot\mathbf{j}$ (where $\mathbf{j}$ is given by Eq. 35) that occurs in the continuity equation derived from the Klein-Gordon equation is obtained from the expression $\psi^*\Delta\psi - \psi\Delta\psi^*$ and that the structure of this expression exhibits a lot of similarity with the structure of the quantity $\psi_1\Delta\psi_2 - \psi_2\Delta\psi_1$ Kirchoff started from in the derivation of his Huyghens' principle. Quantities of this type can thus be used to define mathematical flows that do not need to have a physical meaning.

From the wave equations $\Box\,\psi_1 = 0$ and $\Box\,\psi_2 = 0$ one can also establish a continuity equation for quantities:

$$j_\mu = \imath\,(\psi_1^*\,\partial_\mu\psi_2 - \psi_2\,\partial_\mu\psi_1^*), \tag{37}$$

For $\psi_1 = e^{\imath(Et-\mathbf{p}\cdot\mathbf{r})}$ and $\psi_2 = e^{-\imath(Et-\mathbf{p}\cdot\mathbf{r})}$, we have then:

$$c\rho = \imath(\psi_1^*\,\frac{\partial\psi_2}{\partial ct} - \psi_2\,\frac{\partial\psi_1^*}{\partial ct}) = E(\psi_1^*\psi_2 + \psi_1\psi_2^*). \tag{38}$$

This conundrum disappears if electromagnetic waves are also probability amplitudes, whereby it is now the polarization of the photon which displays the oscillatory behaviour. We grasp our intuition about physical waves from the example of water waves, sound waves or waves propagating along a rope, where the wave transports energy and the energy transport is due to mutual interaction. But probability amplitudes are not physical waves in this intuitive sense. It is only their mathematical expression that makes the probability amplitudes look like waves. Because these probabilities are proportional to numbers of particles and particles represent energy, the probability amplitudes also "transport energy", but the transport mechanism is not based on mutual interaction, just on individual-particle motion.

The quantity $\psi_1^*\psi_2 + \psi_1\psi_2^*$ intervenes in $|\,\psi_1 \boxplus \psi_2\,|^2$ in addition to the positive quantity $\psi_1^*\psi_1 + \psi_2\psi_2^*$, and it can take both positive and negative values leading to constructive and destructive interference. The idea to combine $\psi_1 = e^{\imath(Et-\mathbf{p}\cdot\mathbf{r})}$ and $\psi_2 = e^{-\imath(Et-\mathbf{p}\cdot\mathbf{r})}$ in this argument can be found back in the idea of the paths C_1P and PC_2Q building up the path C_1PC_2Q described above, but it would be more general because it could apply for paths whereby P and Q are both behind the slits. This is certainly necessary if we want to explain interference. In that case the paths C_1B and BC_2 belong to the type of paths that could be considered. The idea of using paths C_1P and PC_2Q only serves to illustrate that $\psi = \psi_1 \boxplus \psi_2$ cannot be an exact rule. The term $\psi_1^*\psi_1 + \psi_2\psi_2^*$ which only occurs when there is interference must thus be due to the backward action of Huyghens' principle, as the presence of complex conjugation shows. This does not imply backward action in physics, because Huyghen's principle is only a purely mathematical tool. But we can relate Huyghens' principle to the fact that we obtain the Dirac equation by extrapolation as pointed out in Footnote 17. Due to this extrapolation, we obtain a Dirac equation for the double-slit experiment. In this equation a path p_1 through slit S_1 must be consistent with a path p_2 through slit S_2. This consistency criterion can be expressed by calculating the phase difference over paths GC_1B and GC_2B, which must be a multiple of 2π. This is exactly the argument we developed in reference [1]. But we can also calculate the phase differences by considering GC_1BC_2G as a loop. Over this loop the phase must be a multiple of 2π. This approach introduces traveling backwards in time. We can also express the consistency criterion by using Huyghens' principle using waves that are traveling backwards in time. In fact, all paths between two points must lead to the same phase difference for global consistency of the wave function. This is thus reason why we can consider all possible paths as Feynman did. All these approaches ensure that the wave function will be a function. One can question the condition that the wave function should be a function. We have given a first justification for it in Subsection 9.1. But we can also justify it is a heuristic method to insure that we treat correctly changes of mass due to e.g. Thomas precession. Precession changes the mass of a particle. Changing the mass will result in a change of orbit, and changing the orbit will change the precession. Therefore we must carefully monitor the precession and the way to do it is to postulate that the wave function must be a function. In certain cases, one even has to replace $\mathbb{R}^3$ by a Riemann manifold (containing multiple copies of $\mathbb{R}^3$) in order to recover a wave function that is really a function. Introducing harmonic polynomials allows then to recover a definition domain that corresponds to $\mathbb{R}^3$ [1].

The rôle played by Eq. 38 in the previous discussion shows that the mathematical flows are not always quantities devoid of any interest. Wave equations should therefore not be dismissed because they do not permit to derive easily an expression for a positive probability density, as was the case for the Klein-Gordon equation. In fact, squaring the free-space Dirac equation automatically yields a free-space Klein-Gordon equation, which has therefore to be correct. The mathematical quantity in Eq. 38 is thus also valid for the Dirac equation.

What all this also shows is that the extrapolation of the wave function from a path to the whole of $\mathbb{R}^4$ we described is all but innocent. As we stated in reference [1]: The Dirac equation is classical. It is the way we solve it that turns it into quantum mechanics! The present analysis confirms this. We can justify the extrapolation procedure by stipulating that we want to obtain an equation for a probability distribution. This emphasizes that the double-slit paradox is a probability paradox. We can perhaps not claim that this analysis is watertight, but it gives a good conceptual grasp on the problem.

11 Strong and weak orthogonality as a criterion for absence or presence of interference

We encountered a few cases where we had $\psi(\mathbf{r}) = \sum_j c_j\psi_j(\mathbf{r})$, whereby $\forall\mathbf{r} \in \mathbb{R}^3 : \psi_j(\mathbf{r})\,\psi_k(\mathbf{r}) = \delta_{jk}[\,\psi_j(\mathbf{r})\,]^2$. This is a much stronger condition than the orthogonality condition we normally use for wave functions and which is $\int_{\mathbb{R}^3} \psi_j^*(\mathbf{r})\,\psi_k(\mathbf{r})\,d\mathbf{r} = \delta_{jk}$. We could call it therefore strong orthogonality. We have then $|\psi(\mathbf{r})|^2 = \sum_j |c_j|^2\,|\psi_j(\mathbf{r})|^2$. This corresponds then to the rule we derived in special cases for the superposition of pure states. But we have seen that for the superposition principle the strong orthogonality does not need to be justified. Strong orthogonality is only needed to avoid interference when we apply Huyghens' principle. When the condition $\forall\mathbf{r} \in \mathbb{R}^3 : \psi_j(\mathbf{r})\,\psi_k(\mathbf{r}) = \delta_{jk}[\,\psi_j(\mathbf{r})\,]^2$ is not satisfied, we run then into the phenomenon of interference. Strong orthogonality implies that the interference term $\psi_j^*(\mathbf{r})\,\psi_k(\mathbf{r}) + \psi_j(\mathbf{r})\,\psi_k^*(\mathbf{r})$ becomes zero over the whole of $\mathbb{R}^3$. From this we can see that the condition $(\forall\mathbf{r} \in \mathbb{R}^3)\,(\psi_j(\mathbf{r})\,\psi_k(\mathbf{r}) = \delta_{jk}[\,\psi_j(\mathbf{r})\,]^2\,)$ as we formulated it is actually too strong, as a strong orthogonality criterion $(\forall\mathbf{r} \in \mathbb{R}^3)\,(\psi_j(\mathbf{r})^*\,\psi_k(\mathbf{r}) + \psi_j(\mathbf{r})\,\psi_k^*(\mathbf{r}) = 2\delta_{jk}\,\psi_j(\mathbf{r})^*\,\psi_j(\mathbf{r})\,)$ would suffice. The latter corresponds to $(\forall\mathbf{r} \in \mathbb{R}^3)\,(\,\mathfrak{R}[\,\psi_j^*(\mathbf{r})\,\psi_k(\mathbf{r})\,] = \delta_{jk}\,\psi_j^*(\mathbf{r})\,\psi_j(\mathbf{r})\,)$. In the Dirac theory this would become $(\forall\mathbf{r} \in \mathbb{R}^3)\,(\psi_j^\dagger(\mathbf{r})\,\psi_k(\mathbf{r}) + \psi_j(\mathbf{r})\,\psi_k^\dagger(\mathbf{r}) = 2\delta_{jk}\,\psi_j^\dagger(\mathbf{r})\,\psi_j(\mathbf{r}))$. This reminds of the orthogonality condition $\Psi_j(\mathbf{r})\,\Psi_k(\mathbf{r}) + \Psi_j(\mathbf{r})\,\Psi_k(\mathbf{r})$ for 4×4 representation matrices of four-vectors in the representation theory of the Lorentz group, especially as the Dirac matrices γ_x, γ_y, γ_z are anti-hermitian, and we are considering here an orthogonality property over $\mathbb{R}^3$ rather than over $\mathbb{R}^4$. Strong orthogonality can in this case easily be obtained, e.g.

$$\begin{pmatrix} \psi_1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ \psi_2 \\ 0 \\ 0 \end{pmatrix} \tag{39}$$

would be strongly orthogonal, whatever the values $\psi_1(\mathbf{r})$ and $\psi_2(\mathbf{r})$ the functions ψ_1 and ψ_2 may take. The functions ψ_1 and ψ_2 themselves do no need to be orthogonal in any sense whatsoever. Strong orthogonality would then easily be obtained in the spinor space and not require orthogonality in function space $F(\mathbb{R}^3, \mathbb{R})$.

This shows the importance of considering wave functions like $\psi_1(\mathbf{r}) \boxplus \psi_2(\mathbf{r})$ as not being a superposition of states. One circumstance that is pointing in this sense is that the result $\psi_1(\mathbf{r}) \boxplus \psi_2(\mathbf{r})$ was shown not to be exact. When we use the weak orthogonality condition $\int_{\mathbb{R}^3} \psi_j(\mathbf{r})^* \, \psi_k(\mathbf{r}) \, d\mathbf{r} = \delta_{jk}$, the quantities ψ_1 and ψ_2 that occur in $\psi_1(\mathbf{r}) \boxplus \psi_2(\mathbf{r})$ may turn out to be fortuitously orthogonal. However, for true probability amplitudes ψ_j we should be able to select any patch $V \subset \mathbb{R}^3$ and the functions ψ_j should behave as probability densities over it. As on such subsets V (imposed on the double-slit experiment) the probabilities would correspond to a true physical problem, we could expect that the wave functions should be orthogonal over V. By considering all possible subsets V, this leads to the idea that the wave functions should satisfy the strong orthogonality condition $\forall \mathbf{r} \in \mathbb{R}^3 : \psi_j(\mathbf{r}) \, \psi_k(\mathbf{r}) = \delta_{jk} [\, \psi_j(\mathbf{r}) \,]^2$ to avoid interference when we apply Huyghens' principle. This leads then to the idea that the isolated functions ψ_1 and ψ_2 in $\psi_1(\mathbf{r}) \boxplus \psi_2(\mathbf{r})$ are not probability densities for the double-slit experiment because they are not orthogonal (unless they are fortuitously orthogonal). This would justify that we are not allowed to calculate $|\psi(\mathbf{r})|^2$ according to $|\psi(\mathbf{r})|^2 = |c_1|^2 |\psi_1(\mathbf{r})|^2 + |c_2|^2 |\psi_2(\mathbf{r})|^2$, because $\psi_1(\mathbf{r}) \boxplus \psi_2(\mathbf{r})$ is not a superposition and explain the origin for the difference between coherent and incoherent summing of probabilities.

12 Also the undecidability does not apply to the single electrons but to the set-up

We must now gather all our results. We have seen that the wave function contains the full information about the experimental set-up. The paths we may explore according to the Huyghens' principle are not the real paths. The wave function does therefore not give access to the real paths, such that we cannot determine them by simulation. The most shocking result is undoubtedly that we do not reproduce the coherent sum rule in the double-slit experiment when we average statistically following binary logic. It would be worth investigating if we could reproduce the coherent sum rule by following ternary logic, because the fundamental paradox can be reproduced in the maths, such that you can scrutinize them to find the answers. They are wrapped up within the subtlety of the Dirichlet problem. Does nature follow ternary logic? And why does nature switch back to binary logic when we put a light source behind the slits as suggested by Feynman? The answer must be that what your experimental results tell you is not that nature follows sometimes ternary logic and sometimes binary logic. The undecidability does not reside in nature itself. What the experimental results tell is if the set-up you designed follows binary logic or ternary logic. It is a verdict about the set-up not about the electrons. The electrons always go through one of the two slits, following binary logic. But it is the logic of the experimental set-up that determines how much you will find out about the true path of the electron. If the set-up is made in such a way that you cannot tell which way the electron went because the interactions have not left the slightest trace of the passage of the electron through the set-up, the logic of your set-up is ternary. But if the set-up is made in such a way that one can tell, then its logic is binary. You can switch between binary and ternary logic at will by modifying the set up. It is the fact that you can switch from one outcome to the other by changing the set-up that strongly suggests that the changes are due to a difference between binary and ternary logic of the set-up, in agreement with Feynman's observations (see Footnote 16). And it is the ease with which you can switch that proves that the electron itself follows binary logic.[30]

If you had divine powers that would permit you to watch the electron without interacting with it, you would have seen and you would know through which slit the electron has traveled. And that would be true all the time. But the problem is that there do not exist set-ups that correspond to this ideal *Gedankenexperiment* of observing the electron without interaction. Your experiments are not a realization of a divine *Gedankenexperiment*. They are done with real-world set-ups, where facts are created by interactions of matter with matter and not by pure thought. Now if you try to average with binary logic over the paths in a set-up that obeys ternary logic, you will contradict yourself by violating the ternary logic of the set-up you designed yourself, and you are fooling yourself with the contradiction you create by ignoring that your set-up can never approach the ideal of watching the electron without interacting with it. You designed a real-world probability problem that must follow ternary logic and then you claim it should follow the ideal-world binary logic. That the difference matters is perhaps astounding but the mathematics already tell it does.[31] We can tentatively interpret the interference fringes as places where the question through which slit the

[30] Adding elements to the set-up K_1 that swap its logic from binary to ternary or *vice versa* implies that we use a modified set-up K_2. Consequently, we will also modify the wave function, which just describes the set-up, from χ_1 to χ_2. We can e.g. assume that the wave functions χ_1 and χ_2 are Fourier transforms of potentials W_1 and W_2. In fact, any change of a set-up K is a change of its potential W and thus of its wave function χ. Set-ups that follow different logics do not have a common distribution function (see Footnote 31). As they correspond to different logics, χ_2 has no bearing on any analysis one may to carry out on the results obtained with K_1, and χ_1 has no bearing on any analysis one may to carry out on the results obtained with K_2. It is especially useless to consider χ_2 as a temporal continuation of χ_1 because that would lead to the wrong over-interpretation that modifying K_1 to K_2 could alter χ_1 by an action that reaches out backwards in time. This is the paradox of the delayed-choice experiment or of the quantum eraser [30], which is due to the choice of a a wrong angle of perspective on the problem, viz. by focusing on the idea that the set-up studies the electrons. The geometry of the set-up is a global and non-local property, which we explore with a large number of electrons to obtain its Fourier transform.

[31] We explained in Subsection 7.3 that this can be understood by pointing out that there does not exist a *common* probability distribution that could be used for the three experiments (with potentials V_1, V_2 and V_3). In fact, the definition domain of the wave function for the double-slit

electron has traveled is decidable (because the opposite phases on the two alternative paths could help us in figuring this out) or undecidable (because the phases on the alternative paths are identical). On the definition domain of an undecidable wave function, the probabilities must become zero in those places where the question could become decidable. These are the places where the interference is destructive, as illustrated in Fig. 3.[32]

We must point out that the path of the electron is determined, but that the fact that you cannot see and collect the corresponding information renders the question which way the electron travelled undecided. One must thus make a clear distinction between determinism and decidability in order to avoid the apparent contradiction that there exist situations where we could call the paths both "determined" and "not determined", by pointing out and defining that these paths are "determined" but effectively "undecidable" within the context of the set-up (rather than "not determined"). This disambiguation will then remove the confusion. Determinism is about the "absolute truth" (Einstein), decidability is about what the set-up of an experiment can decide about that truth (Bohr). There are thus two levels of "truth", an absolute one and an experimental one. The error you make by reasoning with binary logic and the incoherent rule on a set-up that follows ternary logic and a coherent rule is given by the real number $\psi_1^*\psi_2 + \psi_2^*\psi_1$, which expresses the overlap between ψ_1 and ψ_2. The sign of that number can go either way, such that it models the alternating pattern of decidable and undecidable questions. The error can disappear on average over $\mathbb{R}^3$ after integration (weak

experiment (i.e. the potential V_3) when the scattering is coherent does not contain subsets $\mathcal{S}_j$ on which $\psi(\mathbf{r}, t) \neq 0$ and on which one could decide that the electron has traveled through one of the slits S_j. These sets $\mathcal{S}_j$ are just empty and the whole domain of the coherent wave function where $\psi(\mathbf{r}, t) \neq 0$ is undecidable as can be seen from the tentative interpretation of the Fourier transform given immediately hereafter in the main text. In a transition regime between purely classical and purely quantum mechanical the wave function would be $\psi = c_1\psi_1^{(inc)} + c_2\psi_2^{(inc)} + c_3\psi_3^{(coh)}$, with $|c_1|^2 + |c_2|^2 + |c_3|^2 = 1$, where $\psi_1^{(inc)}$, $\psi_2^{(inc)}$, and $\psi_3^{(coh)} \approx \psi_1^{(coh)} \boxplus \psi_2^{(coh)}$ are three different wave functions. They correspond to the answers "yes", "no", "do not know" in ternary logic to the question: "did the electron travel through slit S_1?". The wave function ψ is then a true superposition and we must sum incoherently to obtain the probability: $p = |c_1|^2|\psi_1^{(inc)}|^2 + |c_2|^2|\psi_2^{(inc)}|^2 + |c_3|^2|\psi_3^{(coh)}|^2$. In fact, histories with coherent scattering and with incoherent scattering are incompatible, such that they must be relegated to different wave functions. Here $\psi_3^{(coh)}$ corresponds to coherent scattering, while $\psi_1^{(inc)}$ and $\psi_2^{(inc)}$ correspond to incoherent scattering. In the purely quantum mechanical regime we must put $c_1 = c_2 = 0$ and $c_3 = 1$. In the purely classical regime we must put $c_1 = c_2 = \frac{1}{\sqrt{2}}$, $c_3 = 0$. In the intermediate case we can see very clearly that the coherent probability $|c_3\psi_3^{(coh)}|^2 \approx |c_3(\psi_1^{(coh)} \boxplus \psi_2^{(coh)})|^2$ should not be associated with the incoherent probabilities $|c_1|^2|\psi_1^{(inc)}|^2$ and $|c_2|^2|\psi_2^{(inc)}|^2$. Binary coherent probabilities $|c_1|^2|\psi_1^{(coh)}|^2$ and $|c_2|^2|\psi_2^{(coh)}|^2$ from single-slit set-ups are just not present in the expression $p = |c_1|^2|\psi_1^{(inc)}|^2 + |c_2|^2|\psi_2^{(inc)}|^2 + |c_3|^2|\psi_3^{(coh)}|^2$ for the probabilities in the double-slit set-up. It is only that numerically $\psi_3^{(coh)} \approx \psi_1^{(coh)} \boxplus \psi_2^{(coh)}$ accidentally happens to be a good approximation. This solution in terms of an absence of a *common* probability distribution is very similar to our critique of the derivation of the CHSH inequality in Footnote 22 and in reference [1], p. 278, which shows that this Bell-type inequality cannot be used to analyze the experiments of Aspect *et al.* [17] as has been done to draw the conclusion that quantum mechanics cannot be a local-variable theory (See also reference [31]). In this critique we also argued that there is no proof that there exists a *common* probability distribution for the various measurements that must be carried out to test the CHSH inequality. In absence of such a proof, it is impossible to draw any conclusion from the experiments. For the double-slit experiment, one could adopt a hard line and argue that the number j of the slit S_j through which the particle has traveled is a hidden variable and that it does not exist. But this mixes up the concepts of determinism and decidability as explained further in the main text, and illustrates how a dogma condemning hidden-variable theories could really thwart any further progress in understanding physics in a very detrimental way. We may note that the CHSH inequality is also based on a logic of yes or no answers. We must further note that certain operators in quantum mechanics do not commute and that this entails that they do not have common eigenfunctions. The latter are just probability amplitude distributions. We can thus say that there is no common probability distribution for such non-commuting operators. We see thus that this argument, which comes from quantum mechanics itself, corresponds exactly to our critique of the derivation of the CHSH inequality, which we derived from completely different ideas. There is thus absolutely no reason to pooh-pooh this critique. It permits even to see how we can construct inequalities that will be violated by experiments. When we can answer a question by yes or no within the probability distribution of an operator O_1, that question will thus becomes sometimes undecidable within the probability distribution of an operator O_2 that does not commute with O_1. In the CHSH inequality, there must exist separate probability distributions and eigenfunctions ψ_{jk} for the pairs of photons in the set-up used to measure the quantity $p(A_j \cap B_k)$, that are not common to those in the set-up used to measure the quantity $p(A_{j'} \cap B_{k'})$ with $(j' \neq j) \vee (k' \neq k)$. The inequality can thus not be obtained by integrating over a hypothetical *common* distribution function, because such a distribution function just does not exist. The experiments by Aspect *et al.* are thus very important and very beautiful because they are revealing something very deep and fundamental, but they do not show that "Einstein was wrong". Einstein thought to defeat the conclusions from $[\,A_j, A_k\,] \neq 0$ by considering two correlated photons, and indeed it is possible to measure then $p(A_j \cap B_k)$ as a surrogate for $p(A_j \cap A_k)$, but in the CHSH inequality the commutation relation resurfaces through $[\,A_j, A_{j'}\,] \neq 0 \vee [\,B_k, B_{k'}\,] \neq 0$ when we try to consider $p(A_j \cap B_k)$ and $p(A_{j'} \cap B_{k'})$ simultaneously. What the double-slit experiment shows is that hidden variables do exist but that their probability distributions in one set-up can become useless for the calculation of certain probabilities in an other set-up, because the latter are defined on an incompatible distribution. In a similar spirit, one cannot calculate Malus' law $p(A_j \cap A_k)$ by classical logic from some $p(A_j)$, as any one who has tried may have found out in frustration.

[32] The fact that $\psi \approx \psi_1 \boxplus \psi_2$ can be zero is presumably the best evidence for the fact that this expression must not be interpreted as a physically meaningful sum of two signals ψ_1 and ψ_2, because a real history cannot erase another real history. It is just that the numerical value for ψ is incidentally to a good approximation equal to the numerical result of a semantically completely unrelated, purely mathematical procedure of summing ψ_1 and ψ_2 according to a Huyghens' principle. The only way we found to make sense of the zero values of ψ is the one based on the expression $\psi = \sum_{j \in \mathbb{Z}} c_j \chi_j$ described in Subsection 4.2. What is exact in the extrapolation procedure from $\mathcal{P}$ to $\mathbb{R}^4$ is not the solution $\psi \approx \psi_1 \boxplus \psi_2$ but the wave equation with its boundary conditions expressed in Eq. 16.

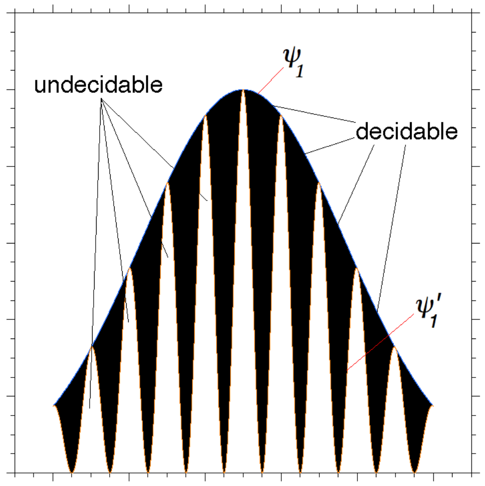

Fig. 3. The contribution of slit S_1 to the probability. The curves display rather the probabilities than the probability amplitudes. The wave function is ψ_1 when only slit S_1 is open (Gaussian curve G) and $\psi_3 = \psi_1 \boxplus \psi_2$ when both slits S_1 and S_2 are open. When we open the second slit, the boundary conditions that define ψ_1 will change. In fact, in certain places $\mathbf{r}$ the phases of ψ_1 and ψ_2 would be opposite. In the traditional approach this corresponds to destructive interference $\psi_3 = \psi_1 \boxplus \psi_2 = 0$, but taking this expression literally entails conceptual difficulties as discussed in Footnote 32. We propose therefore a different reading $\psi_3 = \psi'_1 \boxplus \psi'_2 = 0$, whereby $\psi'_1 = 0$ & $\psi'_2 = 0$, which permits to avoid these difficulties. When the phases of ψ_1 and ψ_2 are different, it would be decidable within the mathematics from which slit the electron has emerged: It would suffice to inspect its phase. Therefore, in order to render ψ_3, ψ'_1 and ψ'_2 entirely undecidable, all the places where $\psi_1 \neq \psi_2$ must be removed from the definition domains of ψ_1, ψ_2 and ψ_3. However, this very sharp picture will be blurred due to the uncertainty $\Delta x \Delta p_x \geq \hbar/2$, (where $\Delta x = D$) about the exact trajectory. Indeed, the wave function ψ_3 must be continuous because it corresponds to the Fourier transform of the potential of the set-up. We obtain thus an interference pattern with a blurred undecidability criterion. In fact, the undecidability requires the wave function ψ_3 to have left-right symmetry and thus to be even in x. It is the procedure to render ψ_3 even that gives it its zeros. We can stipulate that this symmetry argument must be expressed at the centers of the slits, i.e. for $|x| = D/2$. The functions ψ_1 and ψ_2 do not at all comply with the undecidability criterion because they are non-symmetric binary probability amplitudes, and it is meaningless to render their amplitudes G even functions. What we must achieve is that even when the electron truly travels through slit S_1 the experimental knowledge ψ'_1 transmits to ψ_3 must render it undecidable. Therefore we will only render even its phase term. Let us note the phase terms $e^{\imath\phi}$ of the wave functions $Ge^{\imath\phi}$ as χ_j. The even part of $\chi_1(x,t) = e^{-\frac{\imath}{\hbar}(Et-p_x x)}$ is $\chi'_1(x,t) = \cos(Et - p_x x)$. To avoid the conceptual difficulties mentioned, we must imagine that the contribution to ψ_3 from slit S_1 must be the binary logic fulfilling $\psi'_1 = G\chi'_1$ and not ψ_1, such that the modulation of the Gaussian changes from $|\chi_1|^2 = 1$ to $|\chi'_1|^2 = \cos^2(Et - p_x x)$ as shown in the figure. The same reasoning can be applied to ψ_2. The functions ψ'_j will have then the same zeros as ψ_3 and be continuous. Some area shown in black in the figure will be removed from them like in the interference pattern for ψ_3. This leads to $\psi_3 = \psi'_1 \boxplus \psi'_2$ $(= \psi_1 \boxplus \psi_2)$. The result is then that the probabilities for traversing the slits are not the same in binary and in ternary logic. This explains why the classical intuition that we could use an averaging procedure over binary probabilities to take into account the undecidability is flawed and why it does not reproduce the interference: The contributions ψ_1 and ψ'_1 to the wave function from slit S_1 are entirely different. When we "cheat" by acting as though we know that the electron travels through slit S_j and then try to cover up for this by statistical averaging, we do not imagine that $\psi'_j \neq \psi_j$ due to the undecidability. The difference between ψ'_j and ψ_j spoils the whole endeavour. Rather than averaging statistically over the decided probabilities $|\psi_1|^2$ and $|\psi_2|^2$ we should average over the undecided probabilities $|\psi'_1|^2$ and $|\psi'_2|^2$. The precariousness of the whole averaging procedure is further highlighted by the fact that one then still needs an *ad hoc* renormalization of ψ'_1 and ψ'_2 to make sure that the total number of particles is recovered correctly. The result of the highly revised, mixed-logic *ad hoc* procedure becomes then numerically equivalent to the standard calculation $p = |\psi_1 \boxplus \psi_2|^2$ based on the state vector formalism in Hilbert space, whose abstraction impedes any physical understanding of it. This proposal is forced upon us due to the conceptual difficulties mentioned, but it can then also explain why an averaging procedure that has been tried and proved in binary logic fails in ternary logic. It may look contrived, but can be considered as a *reductio ab absurdo* pinpointing the logical flaws in our attempts to reason with binary logic on a situation governed by ternary logic.

orthogonality) and then still show up on a more detailed local scale (absence of strong orthogonality).[33] It is the Fourier transform which is responsible for the uncertainty relations, but these reflect only one possible type of undecidability. In the double-slit

[33] There is certainly information about the actual path of the electron written into the phase difference $\varphi_B - \varphi_G$ an electron builds up over a trajectory GB between source and detector. Within the patch corresponding to χ_j (discussed in Subsection 4.2) the phase difference $\varphi_1 - \varphi_2$ between the two alternative paths GC_1B and GC_2B is $2\pi j$ with $j \in \mathbb{Z}$. The traveling time from source to detector for particles that go through slit S_1 is thus different than that for particles that go through slit S_2. Measuring this time in order to determine through which slit the electron has gone is also subject to an uncertainty relation, and will thus also involve considerations about the interaction of the electrons with parts of the set-up. This problem corresponds to a different set-up with a different wave function ($\psi(\mathbf{r}, t)$ instead of $\psi(\mathbf{r})$).

experiment, it is the rule of thumb $D \lesssim \lambda$ which can be used to predict undecidability with respect to the question through which slit the electron may have traveled. The experimental results honestly reflect the logic of your set-up.

We have seen that the description of the set-up is idealized and cannot account for certain microscopic details. Fig. 1 is a purely geometrical description without any details about atomic positions and interaction potentials and can therefore not possibly be a complete description of the experimental set-up. It is therefore logical to assume that we would not only need hidden variables to deterministically describe the electrons but also to accurately describe the microscopic details of the set-up.

We could ask: “We are not so stupid that we would not be able to reason without contradictions about logic, are we?” Not a single defeatist, not a single quantum priest, not a single editor, not a single peer reviewer, not a single moderator has the right to decide the answer to that question single-handedly and secretly on behalf of whole mankind. Keeping the words of Dieudonné in our mind [10], it is our duty to be proud and confident that the answer will always be “no”. We can appreciate that sometimes it will really take a big fight, because the ambiguity between determinism and decidability was really a hard nut to crack. But even when we have the impression that there is no light, we shall never, ever give up! *“Pour l’honneur de l’esprit humain!”*

13 Synthetic Overview and Conclusion

Because the full argument is complex, an overview of the main ideas is given in the following frames (with interconnections) to be conceived as a flow chart.

❶ Justification of Born’s rule: $p = |\psi|^2$

- $\frac{\partial}{\partial t}$ ⇒ derivation of a continuity equation from the wave equation
- $\frac{\partial^2}{\partial t^2}$ ⇒ (photons): see Feynman
- probability charge-current four-vector must be quadratic expression in terms of spinors

Section 10

❷ Motion of spinning electron on $\mathcal{P}$ ⇒ Dirac-like equation: $\psi = e^{-\frac{i}{\hbar}(Et-\mathbf{p}\cdot\mathbf{r})}\,\psi(0) \rightarrow$ ③
Two extrapolations:

Superposition principle: Dirac-like ⇒ Dirac	Huyghens’ principle: from $\mathcal{P}$ to $\mathbb{R}^4$
introduction of sets, statistical ensembles $\psi = c_1\psi_1 + c_2\psi_2$ ① → $p = \sum_j |c_j|^2|\psi_j|^2$ Solution of the paradox of Schrödinger’s cat Section 2	potentialities, consistent histories $\psi = c_1\psi_1 \boxplus c_2\psi_2$ ① + ③ → $p = |\psi|^2$ Transition: classical mechanics → quantum mechanics Sections 2 - 3, Subsection 7.2, carried out in Section 9 ← ③

Section 1

❸ Extrapolation from $\mathcal{P}$ to $\mathbb{R}^4$ for a simple double-slit potential V (transition CM → QM)

- Fourier transform of V due to spin expressed by $\psi = e^{-\frac{i}{\hbar}(Et-\mathbf{p}\cdot\mathbf{r})}\,\psi(0)$ ← ② (compare with Born approximation)
- proof of Huyghens’ principle: $\psi \approx \psi_1 \boxplus \psi_2$ → ②
- wave function must be a function ⇒ quantization
- ψ contains all the information about the set-up → ④
- Feynman’s path integral is a Huyghens’ principle, the paths are purely mathematical and not physical

Section 9

❹ End of particle-wave duality

- electrons are particles
- probability amplitudes ψ for statistical ensembles of particles are waves (is literally what QM says) ← ①
- ψ is non-local, the electrons interact only locally
- ψ contains all the information about the set-up by Fourier transform as $\mathscr{F}$ is bijective Section 9 ← ③
- Usefull paradigm shift: we are not measuring electrons with a set-up but a set-up with electrons Subsection 6.1

Section 7.1

❺ Undecidability in double-slit experiment Subsection 6.2

- Feynman's analysis hinting at a third possibility (undecided) Sections 4-5
- decided histories ⇒ superposition principle: incoherent summing
- undecided histories ⇒ Huyghens' principle: coherent summing
- Heisenberg's uncertainty relation is one example of a rule of thumb for undecidability
- applying binary averaging to a distribution that follows ternary logic is wrong Subsection 7.3, Section 12
  - ↠ This is due to the absence of a common probability distribution:
    - ↠ is literally what QM says for $p(A_1)$ and $p(A_2)$ when $[A_1, A_2] \neq 0$
    - ↠ compare with criticism of Bell inequalities (derived from commutation relation or differently (see [1], p. 278))
- Another explanation of the interference pattern is possible (see [1], pp. 327-335)

Subsection 7.3

❻ A symmetry description does not provide clues as to the mechanism (see [1], p. 46, pp. 337-340)
↠ agreement with experiment by good fortune
↠ necessity of a case-by-case study (H atom, tunneling)

Various remarks in the text

❼ Strongly orthogonal parts of a wave function ψ do not produce interference

Section 11

In conclusion, we explained how the double-slit paradox can be understood much better by considering it as an experiment whereby one uses electrons to study the set-up rather than an experiment whereby we use a set-up to study the behaviour of electrons. We have also shown that Heisenberg's uncertainty principle is related to Gödels concept of undecidability and how this can be used in an intuitive way to make sense of the paradox of Young's double-slit experiment. Taking this into account leads naturally to the rules that dictate how one should calculate the probabilities for coherent and incoherent scattering in quantum mechanics.

Acknowledgement. G.C. thanks his director, Kees van de Beek for his continuous interest in this work and the CEA for the possibility to carry out this research.

References

1. G. Coddens, in *From Spinors to Quantum Mechanics*, Imperial College Press, (2015).
2. G. Coddens, CEA-01269569 (2016).
3. G. Farmelo, in *The strangest Man. The hidden Life of Paul Dirac: Quantum Genius.* Faber and Faber (2009), p. 430.
4. R.P. Feynman, R. Leighton, and M. Sands, in *The Feynman Lectures on Physics, Vol. 3*, Addison-Wesley, Reading, MA.
5. B.L. van der Waerden in *Group Theory and Quantum Mechanics*, Springer, Berlin-Heidelberg, (1974), translated from *Die gruppentheoretische Methode in der Quantenmechanik*, Springer, Berlin (1932).
6. R.B. Griffiths, in *Consistent Quantum Theory*, Cambridge University Press, Cambridge, (2002).
7. A. Sinha, A.H. Vijay and U. Sinha, Scientific Reports **5**, 10304 (2015); R. Sawant, J. Samuel,A. Sinha, S. Sinha and U. Sinha, Phys. Rev. Lett. **113**, 120406 (2014).
8. K. Gödel, in *"Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I"*, Monatshefte für Mathematik und Physik, **38**,1, pp. 173198 (1931); translated in *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*, Dover, New York, (1992); D.R. Hofstadter in *Gödel, Escher, Bach: An Eternal Golden Braid*, Vintage Books, New York (1989); R.M. Smullyan, in *Gödel's Incompleteness Theorems*, Oxford University Press, New York, (1992).
9. P. Cohen, Proceedings of the National Academy of Sciences of the United States of America **50** (6), 11431148 (1963).
10. J. Dieudonné, in *Pour l'honneur de l'esprit humain - les mathématiques aujourd'hui*, Hachette, Paris (1987).
11. F. Hund, Z. Phys. **117**, 1 (1941); A. Hansen and F. Ravndal, Phys. Scr. **23**, 1033 (1981); P.J.M. Bongaarts and S.N.M. Ruijsenaars, Annals Phys. **101**, 289 (1976).

12. D. Bohm, in *The Undivided Universe: An ontological interpretation of quantum theory,* with B.J. Hiley, Routledge, (1993); D. Bohm, Phys. Rev. **85**, 166-193, (1952).
13. J.A. Wheeler in *Mathematical Foundations of Quantum Theory*, A.R. Marlow editor, Academic Press (1978), pp. 9-48; J.A. Wheeler and W.H. Zurek, in *Quantum Theory and Measurement* (Princeton Series in Physics).
14. A. Tonomura, J. Endo, T. Matsuda and T. Kawasaki, Am. J. Phys. **57**, 117 (1989); see also M.P. Silverman in *More than one Mystery, Explorations in Quantum Interference*, Springer, p. 3 (1995).
15. A.M. Gleason, A. M. (1957). Indiana University Mathematics Journal, **6**, 885 (1957).
16. A. Shimony in *The New Physics*, Paul Davies ed., Cambridge University Press, Cambridge (1983), pp. 373-395.
17. A. Aspect, J. Dalibard, and G. Roger, *Phys. Rev. Lett.* **49**, 91, 1804 (1982); J.F. Clauser and A. Shimony, *Rep. Progr. Phys.* **41**, 1881 (1978); S.J. Freeman and J.F. Clauser, *Phys. Rev. Lett.* **28**, 938 (1972); J.F. Clauser, M.A. Horne, A. Shimony and R.A. Holt, *Phys. Rev. Lett.* **23**, 880 (1969); J.F. Clauser and M.A. Horne, Phys. Rev. D 10, 526 (1974).
18. G. Malinowski, in *Handbook of the History of Logic Volume 8. The Many-Valued and Non-monotonic Turn in Logic*, D.M. Gabbay, J. Woods (eds.), Elsevier, (2009).
19. S. Wagon in *The Banach-Tarski Paradox*, Cambridge University Press, Cambridge, (1985).
20. J. Cramer, Reviews of Modern Physics **58**, pp. 647-688, (1986).
21. T.S. Kuhn, in *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, (1962), see p.114 of 3rd edn.
22. R.S. Longhurst, in *Geometrical and Physical Optics*, Longman, London, (1967).
23. R.P. Feynman, *QED, The Strange Theory of Light and Matter*, Princeton University Press (1988).
24. P.A.M. Dirac, Physikalishe Zeitschrift der Sowjetunion, Band 3, Heft 1 (1933).
25. R.P. Feynman in *The Character of Physical Law*, MIT Press (1967).
26. J.J. Duistermaat in *Huygens' Principle 1690-1990: Theory and Applications*, H. Blok, H. Ferwerds and H.K. Kuiken eds., (Elsevier, 1992), p. 273; J. Hadamard in *Le problème de Cauchy et les Equations aux Dérivés partielles linéaires hyperboliques*, reprint of the 1923 publication, (Dover, New York, 1952).
27. R.L. Wiegel and J.W. Johnson, in *Elements of wave theory*, Proceedings 1st International Conference on Coastal Engineering, Long Beach, California, ASCE, (1950) pp. 5-21.
28. P. Marmier and E. Sheldon, in *Physics of Nuclei and Particles*, Academic Press, New York, (1969).
29. R.P. Feynman, in *The Feynman Lectures on Physics, Vol. 2*, Addison-Wesley, Reading, MA (1970).
30. Y.-H. Kim, R. Yu, S.P. Kulik, Y. Shih et M.O. Scully, Phys. Rev. Lett. **84**, 1-5 (2000).
31. Th.M. Nieuwenhuizen, Foundations of Physics **41**, 580 (2011).