**Noname manuscript No.**
(will be inserted by the editor)

# On information, quanta, and context

**J. Acacio de Barros**

the date of receipt and acceptance should be inserted later

**Abstract** Classical information theory relies on the probabilities that signals appear in a source. As such, information depends on Kolmogorov's probability theory. However, it is well known that some quantum sources are context-dependent, and therefore may not conform to classical probability theory. For such sources, information can be defined through von Neumann's entropy. However, there are possible contextual sources that are neither quantum nor classical. In this paper we explore a simple example of one such source, and show that, similarly to the quantum case, the presence of contextuality increases the amount of information. We also show that, by extending Shannon's entropy to accommodate negative quasi-probability distributions, we are able to provide a measure of informational content that is neither described by Shannon nor von Neumann's entropies.

## 1 Introduction

Information is an fundamental concept not only in engineering, but also in physics. It is present in statistical mechanics, where the thermodynamical behavior of a system can be derived from our lack of information about the underlying states of particles in either a classical or quantum statistics [2]. It is an essential idea in certain interpretations of quantum mechanics [24]. And it is even considered by some physicists as part of the underlying fabric of reality, such as in Wheeler's it-from-bit [47] or D'Arianno's derivation of QFT from information [9].

Information is also key in philosophy of language and philosophy of mind [20]. For example, a very popular view of consciousness is the Integrated Information Theory [38], known as IIT. As its name reflects, IIT relies on the concept of information, and posits that consciousness is a specific type of integration of information where the whole has more information than its parts.

In physics and in engineering, information is usually related to Claude Shannon's definition stemming from his mathematical theory of communications [41]. However, Shannon's theory requires that the information of a set of sources be computed based

School of Humanities and Liberal Studies, San Francisco State University, 1600 Holloway Ave., San Francisco, CA 94132

on the joint probability distribution over the content outcomes of those sources. This creates an issue in physics, where for certain systems no such joint probability exists [23], a phenomenon known as contextuality[1].

Contextual sources appear in physics as a consequence of the impossibility of measuring all observables simultaneously for a quantum system. For instance, Kochen and Specker (KS) [31] proved that for quantum systems whose Hilbert spaces have dimension three or greater, it is possible to construct a set of dichotomic observables (117 in their original paper) and a collection of contexts where some of those observables could be measured simultaneously but where the totality of them could not. They then showed that attempting to assign a truth-value to such observables would lead to an inconsistency, unless the truth-values changed with the context.

This above result from KS implies that we cannot construct a non-negative joint probability distribution for the observables [11], and therefore cannot use Shannon's definition of information. However, for quantum systems, it is possible to define information by using von Neumann's entropy, which is shown to be somewhat equivalent to Shannon's [40]. Furthermore, von Neumann's entropy can be shown relevant to extensions of the quantum case, when there is an orthomodular lattice for the set of observables [28].

In this paper, we investigate the concept of information for contextual sources. In particular, we discuss how Shannon entropy fails to measure the amount of information of contextual sources, and how, for some systems, von Neumann's entropy is not adequate. To do so, we organize this paper the following way. In Section 2 we provide a definition of context and context dependency, also known as contextuality. In Section 3 we show that quantum theory, defined by the algebra of observables on a Hilbert space, requires a description of properties that are context dependent, i.e. that quantum observables are contextual. We do this by following a Kochen-Specker-type argument given in reference [7]. In Section 4 we present Shannon's entropy, considered a measure of information for classical systems, and also von Neumann's entropy, which measures information for quantum systems. Then, in Section 5 we discuss the informational content of several sources that are correlated, some contextual and some non-contextual. We use those examples to show that Shannon's entropy cannot be used in a straightforward way to describe the expected amount of information for contextual systems, a failure that cannot be resolved by resorting to von Neumann's entropy. We end with some final remarks and comments in Section 6.

## 2 Context

In its current and common usage, context comes from linguistics. It means the parts which are related to a particular text, passage, or speech and that are connected to its meaning. For example, the statement "cheap dates are great" has very different meaning if preceded by a conversation about social engagements or by a discussion on the characteristics and benefits of the *Phoenix dactylifera* fruits. One could say that the meaning of the statement changes with the context. But, if we take semantics to be

[1] Some readers may associate the lack of joint probabilities to non-locality, as is the case in Bell's theorem, but the concept of contextuality is more general, with non-locality being a particular case of contextuality where the contexts are selected in a space-like way. See, for example, the discussions about contextuality and non-locality as it relates not only to physics, but also to psychology, in reference [17].

valued with respect to truth values of propositions, one could also say that the truth values changed from context to context: in a given context the statement "cheap dates are great" may be considered by someone as true, whereas in the other context it may be considered by the same person as false.

So, how can we define context based on this idea? This definition would be particularly difficult if we considered that sometimes it is not clear how to assign truth values, particularly as part of speech. For instance, what is the truth value of the statement "Carlos is a good person"? Even if we ignore the ethical assumptions of establishing the truth of such statement, the very definition of "good person" is context dependent, and in a conversation, either with oneself or with someone else, sometimes we are reluctant to provide definite truth values to such propositions. In another paper [13], we argued that a possible general way to deal with context comes from probability theory. In this section, we will present the main idea of representing contexts with probabilities using a theory known as Contextuality by Default (CbD) [21,22][2].

Let us start with the basic ideas. First, we need a probability space, $(\Omega, \mathcal{F}, p)$, where $\Omega$ is a sample space, $\mathcal{F}$ is a $\sigma$-algebra over $\Omega$, and $p$ a function $p : \mathcal{F} \to [0,1]$ satisfying the following properties [32]:

K1. $p(\Omega) = 1$
K2. $p(A \cup B) = p(A) + p(B)$ for $A, B \in \mathcal{F}$ and $A \cap B = \emptyset$.

The probability space $(\Omega, \mathcal{F}, p)$ is just a way to formalize the familiar concept of probability, where the sample space $\Omega$ is simply the set of possible outcomes. For example, if we throw a die, the set of possible outcomes is $\Omega = \{1, 2, \ldots, 5, 6\}$, and if we throw two dice, the set is $\Omega = \{(1,1), (1,2), \ldots, (5,6), (6,6)\}$, where each ordered pair represents the outcomes of the first and second dice, respectively. The algebra $\mathcal{F}$ is a way to talk about more than one outcome: in the case of a die, we can talk about single outcomes, such as $\{1\}$, or we can talk about combinations of outcomes, such as "$\{1\}$ or $\{6\}$" or "not $\{1\}$", and the combination of all the logical propositions we can make over the outcomes of the die would result in an algebra. Finally, $p$ is simply the probability we assign for each element of $\mathcal{F}$. In the case of the die, we have $p(\{1\}) = \ldots = p(\{6\}) = 1/6$, if the die is unbiased. We can immediately compute from these, using K2, that the probability of getting an even number as outcome, $p(\{2,4,6\})$, equals $1/2$.

Now that we have a probability space, we can define a very important tool in using probabilities to describe quantitatively the outcomes of experiments or the truth values of propositions: random variables. Intuitively, random variables are mathematical objects, functions to be precise, that have the same outcomes as an experiment or measurement we wish to model. For example, if we were to model the outcomes of throwing two dice and then summing their values, we could use a random variable that can have as possible outcomes the numbers $1, 2, 3, \ldots, 12$. The random variable corresponding to this example would be a function $\mathbf{R} : \Omega \to \{1, 2, 3, \ldots, 12\}$, where $\Omega = \{(1,1), (1,2), \ldots, (5,6), (6,6)\}$ (all possible outcomes of throwing two dice), defined by $\mathbf{R}((i,j)) = i + j$. Since the probabilities of each element of $\Omega$ is given by $p$, one could use the properties of probabilities to compute the distribution of $\mathbf{R}$, i.e. the probabilities of each outcomes. As importantly, given a data set for an experiment, we can always ask the reverse question: what are the random variables and their corresponding probability space that exhibit the same statistical properties as the data set.

[2] Here we use a simplified version of CbD, which captures some of its main ideas, but is not as precise and general as the formulation presented in the references.

Random variables are important because they provide a consistent way of talking about experimental outcomes, truth values, beliefs, or even about semantics[3]. What we mean is that, by assuming an algebra of events, random variables require that the underlying atomic outcomes (i.e., elements of $\Omega$) provide a consistent description of the processes behind the observed events. To illustrate this, let us look at an example. Imagine we have an experiment where we measure two properties of a system. To keep it simple, let us assume that those properties are binary, i.e., either the system has it or not. Examples of binary properties would be the tossing of a coin (either it is heads or it is not), the passing of an exam (either the grade was sufficient or not), or the success of an offensive play in football (either a goal was scored or not). Let us call those two properties $A$ and $B$. Since they are binary, we can represent the outcomes of those properties by any two numbers, usually $\{0,1\}$ or $\{-1,1\}$, where we can assign, if we wish, 0 or $-1$ to false and 1 to true. The data from observing $A$ and $B$ would give us their individual statistical characteristics, e.g. their expectations, which tells us how often $A$ or $B$ are true, and their joint expectations (if ever measured together), which tells us how $A$ and $B$ are correlated. With such information, we could create a probability space with a sample space $\Omega = \{\omega_{00}, \omega_{01}, \omega_{10}, \omega_{11}\}$ and the following random variables:

$$\mathbf{A}\left(\omega_{00}\right) = \mathbf{A}\left(\omega_{01}\right) = \mathbf{B}\left(\omega_{00}\right) = \mathbf{B}\left(\omega_{10}\right) = 0,$$

$$\mathbf{A}\left(\omega_{10}\right) = \mathbf{A}\left(\omega_{11}\right) = \mathbf{B}\left(\omega_{01}\right) = \mathbf{B}\left(\omega_{11}\right) = 1.$$

Our notation here should be straightforward: $\omega_{00}$ corresponds to $A$ and $B$ being false, $\omega_{10}$ to $A$ true and $B$ false, and so on. If we chose a probability $p$ in a $(\Omega, \mathcal{F}, p)$ that reproduces all the observed expectations and correlations, we created a model of the experimental outcomes in terms of random variables. This model would allow us to infer all possible logical questions about elements of $\Omega$, such as what is the probability of $\omega_{00}$, or "not $\omega_{00}$", or "$\omega_{10}$ and $\omega_{01}$", etc. In other words, all possible logical consequences of the possible values of $A$ and $B$ would be accounted in the statistical model.

Now that we have necessary tools to talk about empirical observations, let us go back to contextuality, first with a simple example. Imagine we have instead of two, three binary properties, namely $X$, $Y$, and $Z$. Let us model then with $\pm 1$-valued random variables, instead of 0 or 1, for symmetry. After several observations, we conclude the following about the statistical properties of the random variables:

$$E\left(\mathbf{X}\right) = E\left(\mathbf{Y}\right) = E\left(\mathbf{Z}\right) = 0, \tag{1}$$

and

$$E\left(\mathbf{XY}\right) = E\left(\mathbf{XZ}\right) = E\left(\mathbf{YZ}\right) = -1. \tag{2}$$

We do not have the expectation of the product $\mathbf{XYZ}$ because we never observe the three variables simultaneously, but only either by themselves or in pairs.

The first set of expectations (1) tell us that each random variable seems to behave randomly, with each outcome $+1$ and $-1$ appearing in our data table with equal probability for each. The second set of expectations (2), involving the second moment, tells us that three random variables are not at all independent. In fact, since the expectation of their products is $-1$, it follows that if one of the variables is $-1$, the other must be $+1$. In other words, if we know $\mathbf{X}$, we also know $\mathbf{Z}$ or $\mathbf{Y}$. However,

[3] See [11,13] for a more detailed discussion of random variables and contextuality as related to those areas.

| $X$ | $Y$ | $Z$ | $XY$ | $XZ$ | $YZ$ |
|---|---|---|---|---|---|
| $-1$ | $-1$ | $-1$ | 1 | 1 | 1 |
| $-1$ | $-1$ | 1 | 1 | $-1$ | $-1$ |
| $-1$ | 1 | $-1$ | $-1$ | 1 | $-1$ |
| $-1$ | 1 | 1 | $-1$ | $-1$ | 1 |
| 1 | $-1$ | $-1$ | $-1$ | $-1$ | 1 |
| 1 | $-1$ | 1 | $-1$ | 1 | $-1$ |
| 1 | 1 | $-1$ | 1 | $-1$ | $-1$ |
| 1 | 1 | 1 | 1 | 1 | 1 |

**Table 1** All possible values of properties $X$, $Y$, and $Z$ and their products.

there is an issue with the above expectations: they lead to a contradiction. To see this, assume that $\mathbf{X} = -1$, which from the first expectation implies that $\mathbf{Y} = \mathbf{Z} = 1$, which contradicts the observation that $\mathbf{YZ} = -1$. A similar contradiction is reached if we assume that $\mathbf{X} = 1$.

The reader may at this point argue that expectations (1) and (2) are an impossibility: one should never have an experiment that leads to contradictions. In a certain sense they would be right, as the contradictions do not come from the expectations, but from the conclusions we are deriving from our model of the expectations. Notice that, as mentioned above, the experimental conditions are such that we never observe all three properties simultaneously, but only two of them. For example, in one observation, properties may be $\mathbf{X} = 1$ and $\mathbf{Y} = -1$, but this tells us nothing about the *unobserved* value of $\mathbf{Z}$. Assuming that the $\mathbf{Z}$ would be the same in this condition as in the other condition leads to the contradiction. In other words, the assumption is that $\mathbf{Z}$ in the context of $\mathbf{X}$ is the same as in the context of $\mathbf{Y}$. This is what we mean by contextuality.

To emphasize this point, consider that, since the binary properties $X$, $Y$, and $Z$, are only observed in pairs, each pair provides a context, as following.

Context 1: $X$ and $Y$
Context 2: $X$ and $Z$
Context 3: $Y$ and $Z$

Since we know those variables are contextual, we model then using random variables that are indexed by their context (see [22] for a more formal procedure). Thus, for Context 1, we have $\mathbf{X}_1$ and $\mathbf{Y}_1$, for Context 2, $\mathbf{X}_2$ and $\mathbf{Z}_2$, and for Context 3, $\mathbf{Y}_3$ and $\mathbf{Z}_3$. It should now be clear that we have no contradiction, as the $\mathbf{X}_1\mathbf{Y}_1 = -1$ tells us nothing about the values of $\mathbf{X}_2$ or $\mathbf{Y}_3$. Contextuality is, in this point of view, manifest by the fact that if we assign for the properties in different contexts the same random variable – i.e. if we impose $\mathbf{X}_1 = \mathbf{X}_2$ , $\mathbf{Y}_1 = \mathbf{Y}_3$ , and $\mathbf{Z}_2 = \mathbf{Z}_3$ – we reach a contradiction.

In the above example, we worked with perfect correlations. But we do not need to have perfect correlations to reach contradictions. For example, consider the assumption that the properties, $X$, $Y$, and $Z$, are *not* contextual. What this means is that, if we were to observe two of those variables, we would be able to assign also a value to the third one in a way that is consistent with the observations. In fact, if we construct a data table with all the possible values, we get something like Table 1. A quick glance at this table shows us that the sum $XY + XZ + YZ$ is never less than $-2$. However, in the contextual example shown above, the sum of the joint pairwise expectations $E(\mathbf{XY}) + E(\mathbf{XZ}) + E(\mathbf{YZ})$ equals $-3$. This observation is at the core of the Suppes-

Zanotti inequalities [44]

$$\begin{aligned} -2 &\leq E(\mathbf{XY}) + E(\mathbf{XZ}) + E(\mathbf{YZ}) \\ &\leq 1 + 2\mathrm{Min}\left\{E(\mathbf{X}), E(\mathbf{Y}), E(\mathbf{Z})\right\}, \end{aligned} \tag{3}$$

which are necessary and sufficient conditions for the non-contextuality of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ when we assume them to be $\pm 1$-value random variables with zero expectation[4].

A useful way to think about the non-contextual bounds of the random variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ is the following. Imagine that we set, for simplicity, the following constraints: $E(\mathbf{X}) = E(\mathbf{Y}) = E(\mathbf{Z}) = 0$, $E(\mathbf{XY}) = E(\mathbf{XZ}) = E(\mathbf{YZ}) = \epsilon$, and $E(\mathbf{XYZ}) = \beta$. Then, it is straightforward to compute that a joint probability exists if and only if $\epsilon$ and $\beta$ are within the bounds of a polytope on the plane $(\epsilon, \beta)$ with the following vertices: $(1, 0)$, $(0, -1)$, $(0, 1)$, and $(-1/3, 0)$. Any correlations outside of this polytope will not allow for a proper joint probability distribution.

It is reasonable at this point to ask whether there are some concrete examples of contextuality. As we mentioned, one could argue that language is contextual [13]. The type of contextuality defined above, in fact, also exists in physics [1,30,39] and in psychology [8]. Here we want to focus on examples from physics, which is what we will do in Section 3.

## 3 Quanta and Context

The world of microscopic phenomena seems to be contextual. But what do physicists mean when they say this? In this section we will investigate the idea of context in physics, and its relationship to the intuitive idea of context expressed in Section 2. Let us start with the famous Kochen-Specker theorem. To understand its origin, let us examine how properties are described in quantum theory.

Quantum theory, in its most basic interpretation, is about how to compute outcomes of experiments. This computation is done the following way. First, we start with the fact that experiments, or, more precisely, measurements done within an experiment, have numerical outcomes (as the length of a rod, or the temperature of an object). Those possible outcomes are represented in quantum theory by real numbers associated to vectors: the numbers are the value of the outcome itself, whereas the vector describes a system with exactly this outcome. We should not think of those vectors as existing in the real three-dimensional space we live in, but in an abstract representational space of physical systems; sometimes this space can be quite large, even infinite-dimensional. The vector space in quantum theory is a Hilbert space, and it has additional properties, such as completeness and metric, but these technical details are not relevant for our purposes.

One of the simplest types of properties, either in quantum physics or outside of it, are binary properties. Those are properties that are either true or false, such as, informally, whether "this brand of ice cream is tasty," "today is cold," or "my car's battery is dead." A physical property can be "the mass of a body is $1.2 \pm 0.1$ kg", and if the body's actual mass lies within the given range, this property is true, otherwise it is false.

Binary properties in quantum theory are represented by projection operators who project any vector in the quantum abstract space to its associated vector. If, for a unit

[4] If the expectations are not zero, then the right hand side of (3) needs to be modified.

vector $\mathbf{w}$ and a projector $\hat{P}$, $\hat{P}\mathbf{w} = \mathbf{w}$, then $\mathbf{w}$ is an eigenvector of $\hat{P}$ with eigenvalue 1, and we can say that $\mathbf{w}$ has the property $P$ associated to $\hat{P}$. For the same projector $\hat{P}$, any unit vector $\mathbf{v}$ on the orthogonal space spanned by $\mathbf{w}$ will have the property $\hat{P}\mathbf{v} = 0$, which means that it is an eigenvector of $\hat{P}$ with eigenvalue 0, and we can say that $\mathbf{v}$ does not have the property $P$. More interestingly, any linear combination $a\mathbf{w} + b\mathbf{v}$, where $a$ and $b$ are complex numbers such that $\sqrt{|a|^2 + |b|^2} = 1$, is not an eigenvector of $\hat{P}$, and we cannot tell whether the system has property $P$ or not if we do not measure it. However, once we measure the system for property $P$, we will either get that $P$ is true or false, and if true the state of the system will "collapse" to $\mathbf{w}$, and if false to $\mathbf{v}$.

For a more concrete example, imagine a four dimensional vector space $\mathbb{C}^4$. The set of vectors $\mathbf{v}_i$, $i = 1, 2, 3, 4$ are a possible orthonormal basis for this space. Given a vector $\mathbf{v}_i$, we can construct an operator $\hat{P}_i$, such that for any other vector $\mathbf{w}$, $\hat{P}_i\mathbf{w} = (\mathbf{w} \cdot \mathbf{v}_i)\,\mathbf{v}_i$, where "$\mathbf{w} \cdot \mathbf{v}_i$" represents the inner product of $\mathbf{w}$ and $\mathbf{v}_i$. Another way to write $\hat{P}_i$ is by using the dual of $\mathbf{v}_i$, which leads to $\hat{P}_i = \mathbf{v}_i\underline{v_i}$, where $\underline{v_i}$ denotes the dual of $\mathbf{v}_i$. In the quantum formalism, each $\mathbf{v}_i$ corresponds to a property: if the physical system we are describing is in a state represented by one of such vectors, then it has the property associated to it, but not to the others. Again, we could also have a system represented by any vector that is a normalized linear combination of the base vectors, i.e. $\sum_i c_i\mathbf{v}_i$. Such vectors do not have any of the associated properties to $\mathbf{v}_i$'s unless we actually measure them, causing them to collapse into one of the basis vectors.

This brings us to an important point. Take the vector $\sum_i c_i\mathbf{v}_i$. There is a projector associated to this vector, and therefore a property as well. And since the space is four dimensional, there are three other properties that are compatible with this vector, such that if the system has its property, it will for sure not have the others. This new projectors correspond to new axis for the splitting of the three dimensional subspace, creating new properties. This idea is behind the proof of the Kochen-Specker theorem [31], which we will present here in a form shown by Cabello et al. [7].

Let us again consider a four dimensional vector space. Following [7] notation, let us consider the projectors $\hat{P}_i$ as indexed with the subscript $i$ corresponding to the 4 components of the vector in a given canonical basis. For example, $\hat{P}_{0,0,0,1}$ corresponds to the projector operator associated to the vector $(0, 0, 0, 1)$. Consider now the following set of equations.

$$\hat{P}_{0,0,0,1} + \hat{P}_{0,0,1,0} + \hat{P}_{1,1,0,0} + \hat{P}_{1,-1,0,0} = 1, \tag{4}$$

$$\hat{P}_{0,0,0,1} + \hat{P}_{0,1,0,0} + \hat{P}_{1,0,1,0} + \hat{P}_{1,0,-1,0} = 1, \tag{5}$$

$$\hat{P}_{1,-1,1,-1} + \hat{P}_{1,-1,-1,1} + \hat{P}_{1,1,0,0} + \hat{P}_{0,0,1,1} = 1, \tag{6}$$

$$\hat{P}_{1,-1,1,-1} + \hat{P}_{1,1,1,1} + \hat{P}_{1,0,-1,0} + \hat{P}_{0,1,0,-1} = 1, \tag{7}$$

$$\hat{P}_{0,0,1,0} + \hat{P}_{0,1,0,0} + \hat{P}_{1,0,0,1} + \hat{P}_{1,0,0,-1} = 1, \tag{8}$$

$$\hat{P}_{1,-1,-1,1} + \hat{P}_{1,1,1,1} + \hat{P}_{1,0,0,-1} + \hat{P}_{0,1,-1,0} = 1, \tag{9}$$

$$\hat{P}_{1,1,-1,1} + \hat{P}_{1,1,1,-1} + \hat{P}_{1,-1,0,0} + \hat{P}_{0,0,1,1} = 1, \tag{10}$$

$$\hat{P}_{1,1,-1,1} + \hat{P}_{-1,1,1,1} + \hat{P}_{1,0,1,0} + \hat{P}_{0,1,0,-1} = 1, \tag{11}$$

$$\hat{P}_{1,1,1,-1} + \hat{P}_{-1,1,1,1} + \hat{P}_{1,0,0,1} + \hat{P}_{0,1,-1,0} = 1. \tag{12}$$

Because all four projectors in each equation (4) to (12) are associated to orthogonal vectors, it follows that the sum of the four projectors is the identity operator. In other words, projecting a vector into four orthogonal components in a four dimensional space, gives us the component in each of those directions, and their sum gives us back the vector we started with. This has a natural interpretation: since each projector corresponds to a possible property of the system, only one of those properties can be true at a time for each line.

If we now assume that quantum properties are not contextual for a quantum system, e.g. the property associated to $\hat{P}_{0,0,0,1}$ is true both in (4) and (5), this leads to a mathematical inconsistency. To see this, simply notice that every $\hat{P}_i$ in (4)–(12) appears twice. Since a property can be either 0 or 1, if follows that if we sum all the lines on the left hand side of (4)–(12) we get an even number. However, if we add the right hand side, we get the number 9, which is odd, a clear contradiction. As discussed in Section 2, this contradiction occurs because we assume the properties are the same in a different context. If we did not assume this, then we would not reach a contradiction. So, quantum observables are contextual.

Let us end this section with a few remarks. First, notice that the contextuality of quantum systems is a consequence of the algebra of observables, defined as projectors in a vector space. This contextuality is state independent, in the sense that it does not matter how you describe your physical system, it will be contextual. However, there are sets of observables that exhibit contextuality only for certain physical systems prepared in special ways [3,4]. This is particularly interesting because, for some especially prepared systems, it is possible to have "contextuality at a distance", in the sense that the value of a property $P$ may depend not only on the local choices of experimental apparatuses, but also on the far away choices of measurement made by other experimenters within a time-like interval. This is why quantum mechanics is though of as a non-local theory.

Our second point is that this type of contextuality seems not only to be unique to quantum systems, but also to bring non-classically reproducible features that may be responsible for the effectiveness of quantum computers [45,29]. For example, it is possible to create a set of inequalities, similar to (3), that shows certain quantum systems as contextual. For those very same systems, it is possible to simulate the quantum correlations that lead to a violation of non-contextuality conditions, even in a non-local way, with classical fields [43]. However, as [34] points out, such violations are not the same as quantum, and they do not imply contextuality, as the quantum systems are made of particles, and not of continuum fields.

Finally, we would like to emphasize that, even though quantum contextuality seems to be special, there exists contextuality outside of quantum physics. As we remarked in Section 2, contextuality is part of linguistics and pragmatics, psychology [8], and it also seems to play an important role in explorations of consciousness [18,35].

In the next section we will investigate the concept of information, and then examine it as it relates to contextuality. We will see that there are some problems associated to contextuality and information that need to be addressed, and we propose a possible route to approach this problem.

## 4 Measuring information

How does one define and measure information? There are many different ways to measure information, but here we focus on the mathematical theory developed by Claude Shannon [41]. Shannon started with the idea that the amount of information a signal carries is related to how surprising that signal is. The more surprising a signal, the more information it carries, whereas the more expected and less surprising a signal is, the less information it carries. For example, the statement "classes will be held as scheduled tomorrow" is less informative than "classes are cancelled for tomorrow", as the latter is the unusual event, whereas the former is so common as to not require, from most students' points of view (and for most instructors), constant reminder and reassurance. Formally, Shannon's definition of information relies on probability theory, considered in some interpretations as a measure of our rational expectations [25].

In his famous paper [41], Shannon used probabilities to define the average amount of information. Let $\mathbf{X}$ be a source, whose output can be any of the events in a set of $N$ possibilities $\{x_1, x_2, \ldots, x_N\}$, and let $p(x_i)$ be the probability that $x_i$ is observed (i.e. the less probable $x_i$ is, the more surprising it should be to see it). Shannon defined the amount of information associated to an observation of $x_i$ as $\log_2 p(x_i)$, and consequently the average amount of information of the source $\mathbf{X}$ would be

$$H(\mathbf{X}) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) . \tag{13}$$

$H(\mathbf{X})$ is known as Shannon's *entropy* of source $\mathbf{X}$.

The use of $\log_2$ in (13) is related to the choice of using as the unit of information a "yes-no" statement that can be coded by a single binary bit, 0 or 1. To understand the intuition behind it, imagine the case where $N = 2$. In this case, we have $\mathbf{X}$ as either $x_1$ and $x_2$, and all we needed is a bit to code which was observed. Now, in the case where $p(x_1) = 0$ and $p(x_2) = 1$ (or, similarly, $p(x_1) = 1$ and $p(x_2) = 0$), there are no surprises: we always get $x_2$ (or $x_1$). Since $\log_2 1 = 0$, it follows that in this case the source's information is zero. The more interesting case is when we have no idea which one, $x_1$ or $x_2$, will occur, and the maximum information we can gain is when both are equiprobable. For the equiprobable case, $p(x_1) = p(x_2) = 1/2$, and $\log_2 1/2 = -1$, which implies that $H(\mathbf{X}) = 1$. We can do the same reasoning for $N = 4$, $N = 8$, $N = 16$, etc, and we will find that when all outcomes are equally probable the entropy is 2, 3, and 4, respectively, which is exactly the number of bits that are required to code each one of the outcomes. In other words, by using the information as $\log_2$ the entropy in (13) gives us the average information in units of bits.

Shannon's information is extremely useful in the cases mentioned above, as well as for engineering applications. For example, one of Shannon's main result was a theorem showing that, for a source $\mathbf{X}$ with entropy $H(\mathbf{X})$, there exists an optimal coding for each outcome of the source that allows a representation of signals from $\mathbf{X}$ with $H(\mathbf{X})$ bits or more. This is a non-trivial result, and has profound consequences in signal transmission, coding, and signal compression.

Despite its success, Shannon's information cannot be used indiscriminately, without some changes, for all types of collections of sources. To see this, let us consider the case of two sources that are perfectly correlated, e.g. a source $\mathbf{S}_1$ whose outcomes are isomorphic to the tossing of a coin, and another source $\mathbf{S}_2$ whose outcomes are exactly the same as $\mathbf{S}_1$, i.e. $\mathbf{S}_1 = \mathbf{S}_2$. It is clear that if we look at $\mathbf{S}_1$, it will seem that its entropy is 1, and the same for $\mathbf{S}_2$, and we can naively conclude that their collective information is greater than 1. However, because they are correlated, this would be a wrong conclusion: combined their information content is 1. So, for a collection of sources the amount of information they generate does not depend on their individual entropy, but on a different measure based on their joint probability. We shall examine this in more detail later, but we emphasize that there are collections of contextual sources that do not have joint probabilities. This is the case, for instance, of some entangled quantum sources [5,23].

So, let us consider now quantum systems. A property $O$ of a system $S$ is represented, in quantum theory, by a Hermitian operator $\hat{O}$ on a Hilbert space $\mathcal{H}$, which are generalizations of the projection operators discussed in Section 3. For this reason, Hermitian operators are called *observables.* The state of the system itself is represented by a vector in $\mathcal{H}$ or by a density operator, defined as an observable that is positive semidefinite and has trace one. States that can be represented by vectors are called pure states, whereas states that must be represented by a density operator (thus having no equivalent in terms of a vector representation) are known as mixed states. For example, for the normalized vector $\mathbf{w} \in \mathcal{H}$ representing a pure state, the density operator associated to it would be $\hat{\rho}_w = \mathbf{w}\underline{\omega}$, where $\underline{\omega}$ is the dual to $\mathbf{w}$. It is easy to prove that a density operator $\hat{\rho}$ is in a pure state if and only if it is idempotent, i.e. $\hat{\rho}\hat{\rho} = \hat{\rho}$, with the implication following from $\hat{\rho}_w$.

While all states represented by a vector in $\mathcal{H}$ are pure states, the density operator allows for other states that are not pure. Consider two linearly independent and normalized vectors, $\mathbf{w}_1$ and $\mathbf{w}_2$. The density operator defined by $\hat{\rho}_M = c_1\mathbf{w}_1\underline{\omega}_1 + c_2\mathbf{w}_2\underline{\omega}_2$ is not idempotent, but it is positive semidefinite with trace one when we choose $c_1 + c_2 = 1$. As such, $\hat{\rho}_M$ has an important physical interpretation: it is how we represent the state of a physical system when we do not know whether it is in the pure state $\mathbf{w}_1$ or $\mathbf{w}_2$, with $c_i$ being the probability of the system to be in state $\mathbf{w}_i$. In other words, the density operator representation allows us to include an uncertainty about the state of the system we are describing, and systems that are not in a pure state (i.e. no uncertainty about the state) are said to have *mixed states.*

Thus, the density operator formalism provides a more general framework to describe quantum systems than the vector formalism: it allows the computation of all quantities available in the vector formalism, plus the same quantities where the state is not known, in a way consistent with classical probability theory. As an example, imagine a three dimensional Hilbert space where a possible basis for it is $\mathbf{e}_i$, $i = 1, 2, 3$. In the vector formalism, the expectation $\left\langle \hat{P}_i \right\rangle = \left| \hat{P}_i \mathbf{w} \right|^2$ of the observable $\hat{P}_i$, where $\hat{P}_i = \mathbf{e}_i \underline{\epsilon}_i$ is a projector, and where $\underline{\epsilon}_i$ is the dual of $\mathbf{e}_i$, give us the probability of observing a system in the pure state $\mathbf{w}$ as having the property $P_i$. This same expectation can be computed by using the expression $\left\langle \hat{P}_i \right\rangle = \mathrm{Tr}\left( \hat{\rho}_w \hat{P}_i \right)$. For the case where we have $\hat{\rho}_M$, as defined above, the linearity of the trace gives us that $\left\langle \hat{P}_i \right\rangle = \mathrm{Tr}\left( \hat{\rho}_M \hat{P}_i \right) = c_1 \mathrm{Tr}\left( \hat{\rho}_1 \hat{P}_i \right) + c_2 \mathrm{Tr}\left( \hat{\rho}_2 \hat{P}_i \right)$, or, equivalently, $\left\langle \hat{P}_i \right\rangle = c_1 \left\langle \hat{P}_i \right\rangle_1 + c_2 \left\langle \hat{P}_i \right\rangle_2$. We can interpret this last expression as saying that the expectation of $\hat{P}_i$ is the weighed expectation for each of the two states composing the mixture. Similarly, for any observable $\hat{O}$, it follows, by the linearity of the trace, that $\left\langle \hat{O} \right\rangle = \mathrm{Tr}\left( \hat{\rho}\hat{O} \right)$.

Let us now go back to the idea of measuring information for quantum systems. The quantum system itself can be thought as being a source of information, equivalent to a source $\mathbf{X}$ in classical communications theory. For a state $\hat{\rho}$, its von Neumann entropy [46] is defined as

$$S = -\mathrm{Tr}\left( \hat{\rho} \log \hat{\rho} \right). \tag{14}$$

To see the relationship between (14) and (13), let us examine the case where we have three binary properties, represented by a three-dimensional Hilbert space with an orthonormal basis $\mathbf{e}_i$, $i = 1, 2, 3$. The observable $\hat{P}_i = \mathbf{e}_i \underline{\epsilon}_i$ have binary outcomes 0 or 1. If the state of the system is given by $\hat{\rho} = \sum_i c_i \mathbf{w}_i \underline{\omega}_i$, it follows that $S = -\sum c_i \log c_i$, which is exactly the same as the Shannon entropy formula if we remember that $c_i$ is a proper probability for having the system with state $\mathbf{w}_i$.

Despite their similarities, there are important differences between quantum and classical entropies. For instance, Shannon's entropy was recovered for orthogonal projection measurements, but if $\hat{\rho}$ was a proper mixture of non orthogonal projectors, this conclusion would not follow. In this paper we will not explore in detail these differences, and the interested reader is referred to [40]. However, we will point out that it is possible to prove an equivalent to Shannon's coding theorem for von Neumann's entropy [40], showing that the similarities between those two measures of information are not coincidental. In fact, in reference [28] the authors argued that for a general orthomodular lattice, which includes more restrictive algebras such as a Boolean one, the natural measure of informational content is von Neumann's entropy, with Shannon's entropy emerging as the measure for classical-like situations.

In this section we presented two ways of measuring information: Shannon and von Neumann entropies. We also discussed the similarities between then, and how we can think of Shannon as a particular case of von Neumann in special cases. In the next section we will explore the concept of context-dependent properties, and how they create a problem for descriptions using classical probability theory. We will use the example in the next section to argue about issues on the definition of information for contextual systems.

## 5 Information and Context

Let us examine the difficulty of defining information for contextual random variables by looking at some examples. Consider two random variables, $\mathbf{X}$ and $\mathbf{Y}$, valued as either $+1$ or $-1$ , and both having equal probability of yielding as outcomes either $+1$ or $-1$. In other words, each of them looks random. If we were to use Shannon's entropy [41] to compute the amount of information $H$ for each one of them separately, we would find that

$$\begin{aligned} H\left(\mathbf{X}\right) &= -p_x \log p_x - p_{\overline{x}} \log p_{\overline{x}} \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \\ H\left(\mathbf{Y}\right) &= 1. \end{aligned} \tag{15}$$

In (15) we are using the simplifying notation that $p_x = p\left(\mathbf{X} = 1\right)$, $p_{\overline{x}} = p\left(\mathbf{X} = -1\right)$, where the line over $x$ means a negative outcome and no line means a positive outcome. The values of $H$ in (15) are telling us that each source has one bit of information, and we might want to conclude that both $\mathbf{X}$ and $\mathbf{Y}$ would have, combined, two bits of information.

However, the conclusion that $\mathbf{X}$ and $\mathbf{Y}$ carries two bits of information may be incorrect. This is because examining each source separately could be misleading. If we have two sources, $\mathbf{X}$ and $\mathbf{Y}$, and they are independent, then the information carried by both sources is simply the sum of each of the sources. But if they are not independent, then their combined information is smaller. For example, if $\mathbf{X}$ and $\mathbf{Y}$ were perfectly correlated, i.e. $E\left(\mathbf{XY}\right) = 1$, then knowing $\mathbf{X}$ implies knowing $\mathbf{Y}$ and vice versa, and the

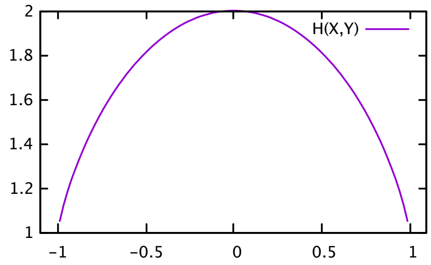

**Fig. 1** Joint entropy of $\mathbf{X}$ and $\mathbf{Y}$ as a function of the correlation $\alpha$.

information contained in the pair would not be two bits, but instead one bit, because of the redundancy. The same conclusion would be arrived if $\mathbf{X}$ and $\mathbf{Y}$ were perfectly anti-correlated, i.e. $E\left(\mathbf{XY}\right) = -1$.

But how can we compute the amount of information from two sources that are neither independent nor perfectly correlated? To do so, we can consider $\mathbf{X}$ and $\mathbf{Y}$ as a pair, and use their joint probabilities. In this case, instead of using as outcomes $-1$ and 1 for each variable separately, we can use all possible outcomes for the pairs, namely $xy$, $x\overline{y}$, $\overline{x}y$, and $\overline{xy}$ (following the notation introduced above). Assuming that $E\left(\mathbf{XY}\right) = \alpha$, it follows that

$$p_{xy} + p_{x\overline{y}} + p_{\overline{x}y} + p_{\overline{xy}} = 1, \tag{16}$$

$$p_{xy} + p_{x\overline{y}} - p_{\overline{x}y} - p_{\overline{xy}} = 0, \tag{17}$$

$$p_{xy} - p_{x\overline{y}} + p_{\overline{x}y} - p_{\overline{xy}} = 0, \tag{18}$$

and

$$p_{xy} - p_{x\overline{y}} - p_{\overline{x}y} + p_{\overline{xy}} = \alpha. \tag{19}$$

Equation (16) is simply the statement that the probabilities of all possible events needs to sum to one (Axiom K1 in Section 2). Equations (17) and (18) are a consequence of the zero expectations for $\mathbf{X}$ and $\mathbf{Y}$, since each outcome $+1$ and $-1$ are equiprobable, whereas (19) is a consequence of $E\left(\mathbf{XY}\right) = \alpha$. The solution to (16)–(19) is

$$p_{xy} = p_{\overline{xy}} = \frac{1}{4}\left(1+\alpha\right), \tag{20}$$

$$p_{\overline{x}y} = p_{\overline{x}y} = \frac{1}{4}\left(1-\alpha\right), \tag{21}$$

and is non-negative for all values of $-1 \leq \alpha \leq 1$. We can use ((20)) and ((21)) to compute the Shannon entropy of those two variables, resulting in

$$H\left(\mathbf{X},\mathbf{Y}\right) = -\frac{1}{4}\left(1+\alpha\right)\log\left(\frac{1}{4}\left(1+\alpha\right)\right) - \frac{1}{4}\left(1-\alpha\right)\log\left(\frac{1}{4}\left(1-\alpha\right)\right).$$

The behavior of $H\left(\mathbf{X},\mathbf{Y}\right)$ is shown in Figure 1. From Figure 1, we can see that if the correlation $\alpha$ is zero, we get maximum information, as expected, whereas the minimum information is when $\mathbf{X}$ and $\mathbf{Y}$ are perfectly correlated ($\alpha = 1$) or anti-correlated ($\alpha = -1$).

Let us now examine the case where we have three ±1-valued random variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ with zero expectation and with correlations $E(\mathbf{XY}) = E(\mathbf{XZ}) = E(\mathbf{YZ}) = \epsilon$. The corresponding equations set by the expectations are

$$p_{xyz} + p_{xy\overline{z}} + p_{x\overline{y}z} + p_{\overline{x}yz} + p_{\overline{xy}z} + p_{\overline{x}y\overline{z}} + p_{x\overline{yz}} + p_{\overline{xyz}} = 1, \tag{22}$$

$$p_{xyz} + p_{xy\overline{z}} + p_{x\overline{y}z} - p_{\overline{x}yz} - p_{\overline{xy}z} - p_{\overline{x}y\overline{z}} + p_{x\overline{yz}} - p_{\overline{xyz}} = 0, \tag{23}$$

$$p_{xyz} + p_{xy\overline{z}} - p_{x\overline{y}z} + p_{\overline{x}yz} - p_{\overline{xy}z} + p_{\overline{x}y\overline{z}} - p_{x\overline{yz}} - p_{\overline{xyz}} = 0, \tag{24}$$

$$p_{xyz} - p_{xy\overline{z}} + p_{x\overline{y}z} + p_{\overline{x}yz} + p_{\overline{xy}z} - p_{\overline{x}y\overline{z}} - p_{x\overline{yz}} - p_{\overline{xyz}} = 0, \tag{25}$$

$$p_{xyz} + p_{xy\overline{z}} - p_{x\overline{y}z} - p_{\overline{x}yz} + p_{\overline{xy}z} - p_{\overline{x}y\overline{z}} - p_{x\overline{yz}} + p_{\overline{xyz}} = \epsilon, \tag{26}$$

$$p_{xyz} - p_{xy\overline{z}} + p_{x\overline{y}z} - p_{\overline{x}yz} - p_{\overline{xy}z} + p_{\overline{x}y\overline{z}} - p_{x\overline{yz}} + p_{\overline{xyz}} = \epsilon, \tag{27}$$

$$p_{xyz} - p_{xy\overline{z}} - p_{x\overline{y}z} + p_{\overline{x}yz} - p_{\overline{xy}z} - p_{\overline{x}y\overline{z}} + p_{x\overline{yz}} + p_{\overline{xyz}} = \epsilon, \tag{28}$$

We now have seven equations set by the correlation and eight variables, which makes this system of equations underdetermined. An additional equation can be set, without loss of generality, by fixing the (sometimes unobservable) triple moment $E(\mathbf{XYZ}) = \beta$, $-1 \leq \beta \leq 1$, which leads to the extra equation

$$p_{xyz} - p_{xy\overline{z}} - p_{x\overline{y}z} - p_{\overline{x}yz} + p_{\overline{xy}z} + p_{\overline{x}y\overline{z}} + p_{x\overline{yz}} - p_{\overline{xyz}} = \epsilon. \tag{29}$$

The solution to equations (22)–(29) is

$$p_{xyz} = \frac{1}{8}(1 + \beta + 3\epsilon), \; p_{xy\overline{z}} = p_{x\overline{y}z} = p_{\overline{x}yz} = \frac{1}{8}(1 - \beta - 3\epsilon) \tag{30}$$

$$p_{\overline{xy}z} = p_{\overline{x}y\overline{z}} = p_{x\overline{yz}} = \frac{1}{8}(1 + \beta - 3\epsilon), \; p_{\overline{xyz}} = \frac{1}{8}(1 - \beta + 3\epsilon), \tag{31}$$

and we can see that some of the $p$'s may be negative for certain values of $\epsilon$ and $\beta$. The non-negative solutions correspond to non-contextual cases, whereas negative solutions are contextual.

Before we compute the entropy based on the joint probability for $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, it is useful to consider some special cases. First, let us examine the perfectly correlated system. In this case, $\epsilon = 1$, and knowing one of the random variables, say $\mathbf{X}$, is sufficient to completely determine all other variables. Therefore, we should expect the entropy for this system to be one bit. The other interesting case is when $\epsilon = 0$, and $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are uncorrelated. At first, it may seem that, because the three variables are uncorrelated, the total entropy for $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ should be three. However, the entropy also depends on the triple moment. For example, if $E(\mathbf{XYZ}) = \pm 1$, even though the variables are pairwise uncorrelated, there is a global correlation between them that reduces the amount of information from three to two bits, since knowing $\mathbf{X}$ and $\mathbf{Y}$ gives us $\mathbf{Z}$. But if, in addition to the pairwise correlations, we have $E(\mathbf{XYZ}) = 0$, then we need three bits[5]. The third interesting example is when $\epsilon = -1/3$, the lower bound of (3). For this case, from (30) and (31) it follows that non-negative solutions exist only if $\beta = 0$, i.e. there is no global correlation, and the triple moment does not give us any extra information. Given that the correlations allow us to explain at most

[5] We point out that a system $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ of random variables where the pairwise expectations are zero and the triple moment is not can be physically realized in quantum mechanics [15].

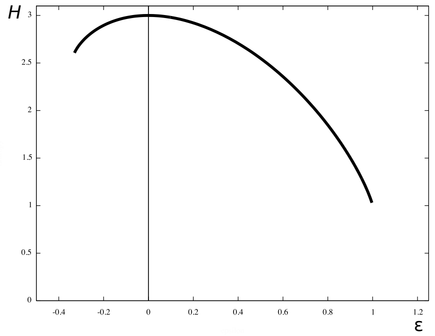

**Fig. 2** Entropy $H(\mathbf{X},\mathbf{Y},\mathbf{Z})$ for $E(\mathbf{XY}) = E(\mathbf{XZ}) = E(\mathbf{YZ}) = \epsilon$ and $E(\mathbf{XYZ}) = 0$ as a function of $\epsilon$. The maximum of 3 bits occurs when $\epsilon = 0$.

1/3 of the other variables, it is reasonable to assume that the information should be more than two bits, but not as high as 3, when all variables are uncorrelated.

A quick examination of Shannon's entropy applied to the probabilities in (30) and (31) shows that our conclusions for $\epsilon = 1$, $\epsilon = 0$ and $\beta = 0$, and $\epsilon = -1/3$ are corroborated, as the entropy for $\epsilon = 1$ is 1, for $\epsilon = \beta = 0$ it hits a maximum at 3 bits, and for $\epsilon = -1/3$ is approximately 2.6 bits. A plot of the entropy $H$ as a function of $\epsilon$ (for the maximal entropy case when $\beta = 0$) is shown in Figure 2. Figure 2 shows the informational content of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ only for the interval $-1/3 \leq \epsilon \leq 1$, as outside of this interval equations (22)–(29) have negative solutions, and Shannon's entropy is not defined. But just because we cannot apply Shannon's entropy, it does not mean that $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ have no informational content outside of the probability polytope. It just means that Shannon's entropy, as currently defined, cannot measure it.

How can we approach this problem? First, let us examine the case when $\epsilon = -1$. A casual analysis could conclude, as in the perfectly correlated example, that the amount of information should be just one bit: if we know $\mathbf{X}$ we know $\mathbf{Y}$ and $\mathbf{Z}$. However, as we saw before $\mathbf{X}$ in the context of $\mathbf{Y}$ cannot be the same as in the context of $\mathbf{Z}$. So, because of this contextuality, the information content should be at least two bits, perhaps more.

To investigate this in more detail, let us imagine that, instead of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, due to the contextuality we have the following random variables: $\mathbf{X_Y}$, $\mathbf{X_Z}$, $\mathbf{Y_X}$, $\mathbf{Y_Z}$, $\mathbf{Z_X}$, and $\mathbf{Z_Y}$. The observed expectations are $E(\mathbf{X_Y}) = E(\mathbf{X_Z}) = E(\mathbf{Y_X}) = E(\mathbf{Y_Z}) = E(\mathbf{Z_X}) = E(\mathbf{Z_Y}) = 0$ and $E(\mathbf{X_Y}\mathbf{Y_X}) = E(\mathbf{X_Z}\mathbf{Z_X}) = E(\mathbf{Y_Z}\mathbf{Z_Y}) = -1$. We can see that the index for each variable indicates their context. For example, $\mathbf{X_Y}$ is the variable $\mathbf{X}$ in the context of being observed with $\mathbf{Y}$. So, we moved from an initial system with three properties to a new system with six. Furthermore, because we have six variables, we would need $2^6 = 64$ equations to uniquely determine the joint probability distribution for those variables. But such equations are unavailable, as we only observe the marginals given by the individual and pairwise expectations.

Because of the correlations, we can reduce the $\mathbf{X_Y}$, $\mathbf{X_Z}$, $\mathbf{Y_X}$, $\mathbf{Y_Z}$, $\mathbf{Z_X}$, and $\mathbf{Z_Y}$ system of random variables from six to three. This suggests that the informational content would be three bits. However, as in the case of three variables, higher moments are not defined. If we impose different values of higher moments, ones that are not consistent with the observed marginals, we could imagine an expansion to the necessary number of random variables, thus increasing the amount of information of the system even beyond 3 bits.

In other words, if we were to draw from the case of two variables, it is not unreasonable to expect that the maximum value for the entropy would be achieved when the higher moments were less correlated. Thus, highly pair-wise correlated sources that are contextual may have more information than non-contextual correlated ones. In fact, a highly correlated contextual set of sources can have more information than the same number of sources when completely uncorrelated.

### 5.1 Information and Negative Probabilities

We now end this discussion with a possible description of contextual information in terms of non-standard probabilities. For the above three-variable example, we may always construct a joint pseudo-probability distribution that is consistent with the observed expectations. This can be done either with upper and lower probabilities (as in [42,16,27]) or with negative probabilities (see [10,14,11]). Here we will use negative probabilities, as they provide a more convenient computational tool than upper probabilities, in particular for quantum physics (see [36,37]).

"Negative probabilities" were introduced by Dirac in his famous Bakerian Lectures [19], in the context of quantum field theory. Though the name *negative probabilities* may be misleading, as only some probabilities of atomic events may be negative, but no observable probabilities may be negative, this terminology is widely used in physics.

Let $(\Omega, \mathcal{F}, p^*)$ be an extended probability space, defined as satisfying axioms K1 and K2 (see Section (2)) but with $p^* : \mathcal{F} \to \mathbb{R}$. We call $p^*$ negative probabilities, as it satisfies the same axioms as Kolmogorov probabilities, except that the non-negativity requirement is relaxed, which means that probabilities measures of the atomic elements can take negative values. A more formal definition of negative probabilities, one that takes into account only the observable marginal probabilities, can be found in reference [11], and the interested reader is referred to it.

As above mentioned, the term, negative probabilities, may be misleading, for the following reasons. First, as mentioned, negative probabilities are not necessarily negative. It is usually the case that, for all atomic elements in $\Omega$, some probabilities are non-negative and others are negative. A quick examination of ((30))-((31)) shows this to be the case for $\epsilon = -1$. Second, it is never possible to observe a negative probability. Though for $\omega_i \in \Omega$ it is possible to have $p^*(\omega_i) < 0$, the observable elements of $\mathcal{F}$, corresponding to the experimental random variables, never have negative probabilities. In fact, the Boolean sub-algebras of observables are always compatible with a proper probability measure. So, even though the term used in physics is *negative probabilities*, it may be more appropriate to call them *signed probabilities*, as those mathematical objects are known in measure theory [26]. Here we will follow the physics tradition and call $p^*$ negative probabilities.

Let us use negative probabilities to define an extension of Shannon's entropy for contextual systems. We start with the extended entropy $S_{NP}$ for $p^*$ as given by the following expression[6],

$$S_{NP} = -\sum_{\omega_i \in \Omega} \left| p^*(\omega_i) \right| \log_2 \left| p^*(\omega_i) \right| . \tag{32}$$

[6] The idea of using something similar to ((32)) was first suggested informally to the author by Patrick Suppes on the context of quantum information, but he did not pursue this idea.

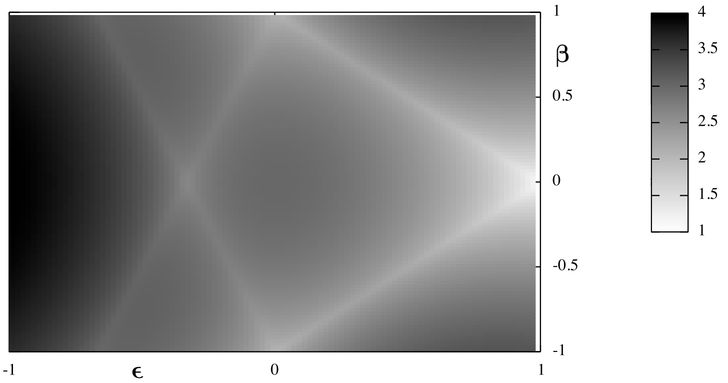

**Fig. 3** Surface plot of $S_{NP}$ as a function of $\epsilon$ (horizontal axis) and $\beta$ (vertical axis). Lighter regions correspond to less entropy, whereas darker regions to more entropy.

The idea of using the absolute value of the negative probabilities comes from using the expression $A = \sum_{\omega_i \in \Omega} |p^*(\omega_i)|$ as a measure of how contextual a probabilistic system is [10]. If $A = 1$, this means that there is no contextuality, since $A = \sum_{\omega_i \in \Omega} p^*(\omega_i)$, which implies that $p^*$'s are non-negative. However, if $A > 1$, then the system is contextual, and the further $A$ departs from one, the stronger the contextual relations between the variables. So, it is not unreasonable to use $|p^*(\omega_i)|$ as a measure of how "surprising" some outcome is, given its close connections to proper probabilities.

From equation ((32)) we can see that $S_{NP}$ is exactly the same as Shannon's entropy for cases when $p^*$ is non-negative. In other words, when a proper joint probability distribution exists, $S_{NP} = H$. It should also be clear that, whenever we have contextuality, the quantity $\sum |p^*(\omega_i)|$ is greater than one, which means that the entropy may increase, as expected from our examples.

In Figure (3) we show a graph of three ±1-valued random variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ with zero expectation, correlations $E(\mathbf{XY}) = E(\mathbf{XZ}) = E(\mathbf{YZ}) = \epsilon$, and triple moment $E(\mathbf{XYZ}) = \beta$. There are several characteristics of this graph that are interesting. First, we can clearly see the non-contextual polytope as the quadrilateral defined by the vertices $(-\frac{1}{3}, 0)$, $(0, 1)$, $(0, -1)$, $(1, 0)$, as the boundary has lower entropy, as expected. We can also see from the figure that the minimum value for the entropy is 1, which occurs when $\epsilon = 1$ and $\beta = 0$, as discussed above. Still within the classical bounds, the region shown has maximum entropy when $\epsilon = \beta = 0$, consistent with our analysis above. More interestingly, as can be seen from Figure (4), the maximum entropy is reached when $\epsilon = -1$ and $\beta = 0$, as was predicted by our random variable argument. So, it seems that the entropy with negative probabilities, as defined by ((32)), may provide a consistent measure of informational content for contextual sources.

## 6 Final Remarks

In this paper we discussed issues related to measuring the informational content of contextual sources. This is an important subject, as it has been argued, for example, that contextuality is a key resource for quantum computation [45,29]. But, more inter-

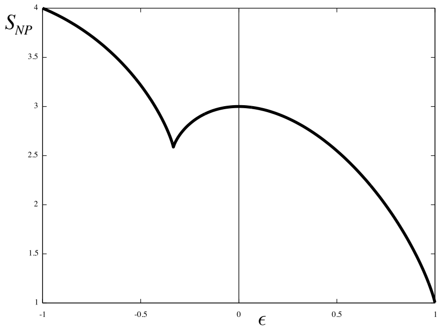

**Fig. 4** Cross section of $S_{NP}$ as a function of $\epsilon$ and for $\beta = 0$. We see a local maxima at $\epsilon = 0$ and a global maxima at $\epsilon = -1$, where $S_{NP} = 4$.

estingly, contextuality is a key concept in quantum physics, and perhaps what makes quantum mechanics different from classical mechanics [6].

In our discussion, we provided a simple example of a contextual set of sources, and we showed that this set of sources may provide more information when it is contextual then when it is not. For instance, when two of the three sources are perfectly anti-correlated, this correlation reduces their informational content to at most two bits, if they are non-contextual. However, if they are contextual, the informational content of the three variables may be as high as three bits. More interestingly, the maximum informational content that three non-contextual sources can have is 3 bits, whereas, as we argued above, contextual sources can have upwards of 4 bits. In other words, contextuality leads to more information. This result should not be surprising. If we think about the meaning of contextuality, it is exactly that a source depends on the context; what we thought were the same sources in two different contexts (or more) were actually two different (and perhaps uncorrelated) sources.

We also showed that we can provide a measure of informational content by extending Shannon's entropy to negative probabilities. This new entropy reduces to the classical Shannon entropy for systems with a proper joint probability distribution, but can be used for regions outside of the classical probability polytope. Our analysis showed that the entropy $S_{NP}$ based on negative probabilities give results that are consistent with a simplified contextuality-by-default description of the sources.

We believe our result unveils interesting applications of negative probabilities to the study of contextual information, not only of quantum information. Furthermore, it opens up several questions that are worth exploring. For example, if $S_{NP}$ is a good representation of the informational content of a system, then it should be possible to prove a Shannon coding-type theorem for negative probabilities. Also, it would be interesting to explore the relationship between $S_{NP}$ and von Neumann's entropy, which we know is related to Shannon's, but can also be applied to certain contextual systems. Finally, as we mentioned, von Neumann's entropy cannot be used in situations such as our three random-variable example, as in a Hilbert space description three pairwise commuting observables have a common base, and therefore a joint probability distribution. Therefore, it is an open question whether $S_{NP}$ can be used for applications outside of physics, such as in psychology, linguistics analysis, or consciousness studies. For example, in consciousness studies a popular approach, known as Integrated Information Theory [38], faces issues with contextuality [12,35].

We emphasize that negative probabilities do not provide the most general treatment for measuring information of contextual sources. For example, negative probabilities require that random variables in different contexts are consistently connected (what is called in physics “no-signaling”) [36]. A more general approach should probably come from an expanded set of contextual random variables, as it is proposed in Contextuality by Default (CbD). CbD can deal with random variables whose mean values vary from context to context (i.e. they are inconsistently connected) [33]. However, because of the extra variables, it is not clear how to deal with information for contextual variables within this approach, and probably a strategy using CbD will need to impose symmetries or utilize minimization principles. Accordingly, negative probabilities reduce significantly the number of required equations for describing contextual sources, as compared to CbD, thus imposing lots of additional symmetries that may not be satisfied by certain systems. The connection between information with negative probabilities and information using CbD, as well as the limitations of negative probabilities, are topics that needs to be explored further.

**Acknowledgements** This paper benefited from conversations with Carlos Montemayor, Gary Oas, Pawel Kurzynski, Federico Holik, and Gregg Jaeger, to whom I am indebted. I would also like to thank the editors of this book, Francisco Doria and Shyam Wuppuluri, for their encouragement and support. It is a great pleasure to dedicate this paper to Professor Chaitin, and it is an honor to have it appear on his Festschrift.

## References

1. Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental Tests of Realistic Local Theories via Bell’s Theorem. *Physical Review Letters*, 47(7):460–463, August 1981.
2. Roger Balian. *Du microscopique au macroscopique: cours de physique statistique de l’Ecole Polytechnique*. Ecole Polytech. Palaiseau, Palaiseau, 1982.
3. J.S. Bell. On the Einstein-Podolsky-Rosen paradox. *Physics*, 1(3):195–200, 1964.
4. J.S. Bell. On the Problem of Hidden Variables in Quantum Mechanics. *Rev. Mod. Phys.*, 38(3):447–452, 1966.
5. Samuel L. Braunstein and Carlton M. Caves. Information-Theoretic Bell Inequalities. *Physical Review Letters*, 61(6):662–665, August 1988.
6. A. Cabello. Quantum physics: Correlations without parts. *Nature*, 474(7352):456–458, June 2011.
7. A. Cabello, J.M. Estebaranz, and G.C. Alcaine. Bell-Kochen-Specker theorem: A proof with 18 vectors. *Physics Letters A*, 212(4):183–187, March 1996. arXiv:quant-ph/9706009.
8. Víctor H. Cervantes and Ehtibar N. Dzhafarov. Snow queen is evil and beautiful: Experimental evidence for probabilistic contextuality in human choices. *Decision*, 5(3):193–204, 2018.
9. Giulio Chiribella, Giacomo Mauro D’Ariano, and Paolo Perinotti. Informational derivation of quantum theory. *Physical Review A*, 84(1):012311, July 2011.
10. J. Acacio de Barros, Ehtibar N. Dzhafarov, Janne V. Kujala, and Gary Oas. Measuring Observable Quantum Contextuality. In Harald Atmanspacher, Thomas Filk, and Emmanuel Pothos, editors, *Quantum Interaction*, number 9535 in Lecture Notes in Computer Science, pages 36–47. Springer International Publishing, July 2015.
11. J. Acacio de Barros, Janne V. Kujala, and Gary Oas. Negative probabilities and contextuality. *Journal of Mathematical Psychology*, 74:34–45, October 2016.
12. J. Acacio de Barros, Carlos Montemayor, and Leonardo P. G. De Assis. Contextualit in the Integrated Information Theory. In *Quantum Interaction*, Lecture Notes in Computer Science, pages 57–70. Springer, Cham, 2017.
13. J. Acacio de Barros, Carlos Montemayor, Leonardo P. G. De Assis, Paul Skokowski, and John Perry. On a Mathematical Representation of Linguistic Contextuality. *Submitted.*, 2018.

14. J. Acacio de Barros, Gary Oas, and Patrick Suppes. Negative probabilities and Counterfactual Reasoning on the double-slit Experiment. In J.-Y. Beziau, D. Krause, and J.B. Arenhart, editors, *Conceptual Clarification: Tributes to Patrick Suppes (1992-2014)*. College Publications, London, 2015.
15. J. Acacio de Barros and P. Suppes. Strict Holism in a Quantum Superposition of Macroscopic States. *Arxiv preprint quant-ph/0003046; to appear in the South American Journal of Logic*, 2000.
16. J. Acacio de Barros and P. Suppes. Probabilistic Inequalities and Upper Probabilities in Quantum Mechanical Entanglement. *Manuscrito*, 33(1):55–71, 2010.
17. J. Acacio de Barros and Patrick Suppes. Quantum mechanics, interference, and the brain. *Journal of Mathematical Psychology*, 53(5):306–313, October 2009.
18. José Acacio de Barros, Federico Holik, and Décio Krause. Contextuality and Indistinguishability. *Entropy*, 19(9):435, August 2017.
19. P.A.M. Dirac. Bakerian Lecture. The Physical Interpretation of Quantum Mechanics. *Proceedings of the Royal Society of London B*, A180:1–40, 1942.
20. Fred Dretske. *Knowledge and the Flow of Information*. MIT Press, Cambridge, Massachusetts, 1981.
21. Ehtibar N. Dzhafarov and Janne V. Kujala. Context-Content Systems of Random Variables: The Contextuality-by-Default Theory. *arXiv:1511.03516 [quant-ph]*, November 2015. arXiv: 1511.03516.
22. Ehtibar N. Dzhafarov and Janne V. Kujala. Contextuality-by-Default 2.0: Systems with Binary Random Variables. In J. Acacio de Barros, Bob Coecke, and Emmanuel Pothos, editors, *Quantum Interaction: 10th International Conference, QI 2016*, volume 10106 of *Lecture Notes in Computer Science*. Springer International Publishing, 2017. arXiv: 1604.04799.
23. A. Fine. Hidden Variables, Joint Probability, and the Bell Inequalities. *Physical Review Letters*, 48(5):291–295, February 1982.
24. Christopher A. Fuchs and Rudiger Schack. Quantum-Bayesian coherence. *Reviews of Modern Physics*, 85(4):1693–1715, December 2013.
25. M. C Galavotti. *Philosophical introduction to probability*, volume 167 of *CSLI Lecture Notes*. CSLI Publications, Stanford, CA, 2005.
26. P.R. Halmos. *Measure Theory*. Springer-Verlag, New York, NY, 1974.
27. Stephan Hartmann and Patrick Suppes. Entanglement, Upper Probabilities and Decoherence in Quantum Mechanics. In Mauricio Suárez, Mauro Dorato, and Miklós Rédei, editors, *EPSA Philosophical Issues in the Sciences*, pages 93–103. Springer Netherlands, January 2010.
28. Federico Holik, A. Plastino, and Manuel Sáenz. Natural Information Measures in Cox's Approach for Contextual Probabilistic Theories. *Quantum Information and Computation*, 16(1&2):0115–0133, 2016.
29. Mark Howard, Joel Wallman, Victor Veitch, and Joseph Emerson. Contextuality supplies the 'magic' for quantum computation. *Nature*, 510(7505):351–355, June 2014.
30. G. Kirchmair, F. Zahringer, R. Gerritsma, M. Kleinmann, O. Guhne, A. Cabello, R. Blatt, and C. F. Roos. State-independent experimental test of quantum contextuality. *Nature*, 460(7254):494–497, July 2009.
31. Simon Kochen and E. P. Specker. The Problem of Hidden Variables in Quantum Mechanics. *Journal of Mathematics and Mechanics*, 17:59–87, 1967.
32. A.N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Co., Oxford, England, 2nd edition, 1956.
33. Janne V. Kujala, Ehtibar N. Dzhafarov, and Jan-Ake Larsson. Necessary and Sufficient Conditions for an Extended Noncontextuality in a Broad Class of Quantum Mechanical Systems. *Physical Review Letters*, 115(15):150401, October 2015.
34. Marcin Markiewicz, Dagomir Kaszlikowski, P. Kurzynski, and Antoni Wojcik. From contextuality of a single photon to realism of an electromagnetic wave. *npj Quantum Information*, 5(1):5, January 2019.
35. C. Montemayor, J. A. de Barros, and L. P. G. De Assis. Implementation, Formalization, and Representation: Challenges for Integrated Information Theory. *Journal of Consciousness Studies*, 26(1):107–132, 2019.
36. G. Oas, J. Acacio de Barros, and C. Carvalhaes. Exploring non-signalling polytopes with negative probability. *Physica Scripta*, T163:014034, 2014.
37. Gary Oas and J. Acacio de Barros. A Survey of Physical Principles Attempting to Define Quantum Mechanics. In Ehtibar Dzhafarov, Ru Zhang, and Scott M. Jordan, editors, *Contextuality From Quantum Physics to Psychology*. World Scientific, 2015.

38. Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol*, 10(5):e1003588, May 2014.
39. M. H. M. Passos, W. F. Balthazar, J. Acacio de Barros, C. E. R. Souza, A. Z. Khoury, and J. A. O. Huguenin. Classical analog of quantum contextuality in spin-orbit laser modes. *Physical Review A*, 98(6):062116, December 2018.
40. Benjamin Schumacher. Quantum coding. *Physical Review A*, 51(4):2738–2747, April 1995.
41. Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.
42. P. Suppes and M. Zanotti. Existence of hidden variables having only upper probabilities. *Foundations of Physics*, 21(12):1479–1499, 1991.
43. Patrick Suppes, J. Acacio de Barros, and Adonai Sant'Anna. Violation of Bell's inequalities with a local theory of photons. *Foundations of Physics Letters*, 9(6):551–560, 1996.
44. Patrick Suppes and Mario Zanotti. When are probabilistic explanations possible? *Synthese*, 48(2):191–199, 1981.
45. Victor Veitch, Christopher Ferrie, David Gross, and Joseph Emerson. Negative quasi-probability as a resource for quantum computation. *New Journal of Physics*, 14(11):113011, November 2012.
46. J. von Neumann. *Mathematical foundations of quantum mechanics*. Princeton University Press, Princeton, NJ, translated by robert t. beyer from the 1932 german edition, 1983.
47. John A. Wheeler. Information, physics, quantum: The search for links. In W. H Zurek, editor, *Complexity, entropy, and the physics of information*, volume 8. CRC Press, Boca Raton, FL, 1990.