arXiv:2406.00717v2 [quant-ph] 31 Mar 2026

Resource-Theoretic Hierarchy of Contextuality for General Probabilistic Theories

Lorenzo Catani[1], Thomas D. Galley[2,3], and Tomáš Gonda[4]

[1]International Iberian Nanotechnology Laboratory, Av. Mestre José Veiga s/n, 4715-330 Braga, Portugal
[2]Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Boltzmanngasse 3, A-1090 Vienna, Austria
[3]Vienna Center for Quantum Science and Technology (VCQ), Faculty of Physics, University of Vienna, Vienna, Austria
[4]Institute for Theoretical Physics, University of Innsbruck, Austria

In this work we present a hierarchy of generalized contextuality. It refines the traditional binary distinction between contextual and noncontextual theories, and facilitates their comparison based on how contextual they are. Our approach focuses on the contextuality of prepare-and-measure scenarios, described by general probabilistic theories (GPTs). To motivate the hierarchy, we define it as the resource ordering of a novel resource theory of GPT-contextuality. The building blocks of its free operations are classical systems and univalent simulations between GPTs. These simulations preserve operational equivalences and thus cannot generate contextuality. Noncontextual theories can be recovered as least elements in the hierarchy. We then define a new contextuality monotone, called classical excess, given by the minimal error of embedding a GPT within an infinite classical system. In addition, we show that the optimal success probability in the parity oblivious multiplexing game also defines a monotone in our resource theory. Finally, we discuss whether the non-free operations can be understood as implementing information erasure and thus explaining the fine-tuning aspect of contextuality.

Lorenzo Catani: lorenzo.catani4@gmail.com
Thomas D. Galley: thomas.galley@oeaw.ac.at
Tomáš Gonda: tomas.gonda@uibk.ac.at

1 Introduction

Generalized (non)contextuality. A crucial research question in the foundations of quantum theory is to identify those features of quantum theory that constitute a true departure from the classical worldview. Addressing this question requires one to first establish a good notion of classicality, which adequately captures the classical worldview. We believe that a good notion of classicality should satisfy the following desiderata (see also [1]): (1) it endorses a principle that defines a clear boundary between aspects that pose interpretational issues and those that do not, (2) it is of broad range of applicability, (3) it is empirically testable, and (4) its violation constitutes a resource for practical applications, in particular in quantum information processing.[1] Motivated by these desiderata, a leading notion of classicality is *generalized noncontextuality* [2].[2] A noncontextual theory is one that is compatible with a classical realist explanation of its operational predictions — it admits of a *noncontextual ontological model*. In such a model, any two experimental procedures that the theory predicts to be operationally indistinguishable also have the same ontological representation (see Section 2 for a precise definition).

Noncontextuality satisfies (1), in that it is an instance of a methodological principle inspired by Leibniz's principle of the identity of indiscernibles [3] (also formulated as a no fine-tuning principle [4]). A violation of such principle would indeed entail an interpretational problem, as it would attribute a conspiratorial connotation to the realist explanation of the theory: why should experimental procedures predicted by the theory to be indistinguishable *in principle* be represented by different distributions in the ontological model? Noncontextuality also satisfies (2), as it applies to a wide range of scenarios including prepare-and-measure experiments of a single system, unlike Bell's local causality — another leading notion of classicality. Moreover, in situations where these are applicable, it coincides with notions of classicality such as non-negativity of quasi-probability representations [5, 6] and Bell's local causality [7]. It satisfies (3), as witnessed by the experiments performed to test quantum violations of generalized noncontextuality [8, 9]. Finally, desideratum (4) has been argued by the numerous works showing that contextuality is a resource for information processing tasks [10–21].

Hierarchy of contextuality. In relation to the last point, contextuality can be used to witness and characterize the advantage of using *quantum* physical systems in practical applications. In such contexts, it is important to know not just whether a theory is contextual, but also to *quantify* how contextual it is. Our article aims to address this by introducing a *hierarchy of contextuality*, in which not all contextual phenomena are equivalent. In particular, it allows us to make more fine-grained distinctions and to quantify the amount of contextuality present in a theory.

To motivate the hierarchy we propose, we take inspiration from the framework of resource theories [22–24]. There, a resource object r (such as a quantum state) is deemed more valuable than a resource object s, denoted $s \preceq r$, if there is a *free operation* that transforms r to s. The preorder relation[3] $\preceq$ is called the *resource ordering*. A paradigmatic example is that of entanglement ordering of bipartite quantum states, where the free channels consist of local operations and classical communication [25]. In this sense, a resource theory (and its associated resource ordering) is defined by a choice of resource objects and of free operations [26]. For instance, if the resource objects studied are quantum states, then the set of free operations is typically a subset of quantum channels closed under composition.

In our case, resource objects are instead physical theories, since (non)contextuality is a property of theories rather than states. However, as contextuality can be witnessed by individual systems within the theory, we restrict our investigation to individual systems. That is, we identify each resource object as a system in a general probabilistic theory (GPT) [27–30], such as classical probability theory, quantum theory, or a subtheory thereof like the stabilizer subtheory [31]. The GPT

[1] This last desideratum is motivated by the belief that identifying the true nonclassicality of quantum theory will ultimately provide the answer to the question about the origin of the alleged quantum computational speed-up.

[2] In what follows we will often omit "generalized" and just talk of "(non)contextuality".

[3] A preorder is a relation that is both reflexive and transitive. It is a partial order if it is also anti-symmetric. If, in addition, any two elements are related, then it is a total order.

system specifies all possible probabilistic behaviours of this physical system and thus characterizes its information-theoretic properties.

Since our resources are not states of a physical system, but rather physical theories, this implies that the transformations that we consider cannot be standard physical operations, as is the case for traditional resource theories of quantum states. Our operations of interest are *simulations*, which are transformations between theories that faithfully encode the information of one theory within another. Simulations that preserve indistinguishability have been introduced in [32] as *univalent* simulations. Since noncontextuality of an ontological model means that the model corresponds to a univalent simulation, one can show [32] that a GPT system is noncontextual if and only if it can be simulated in a univalent way by a classical GPT system — one, whose states form a simplex of probability distributions on a finite sample space. Classical GPT systems thus cannot generate contextuality under univalent simulations. Therefore, we propose the following hierarchy of (generalized) contextuality:

> A GPT system B is said to be **at least as contextual** as a GPT system A if there exists a univalent simulation of A by a composite of B and a classical GPT system.

It is the resource ordering of a resource theory with objects given by GPT systems and free operations given by univalent simulations with free access to classical systems. In Section 5.1, we discuss a possible interpretation of (non-univalent) simulations as expressing a particular kind of information erasure.

Contextuality monotones. As is common in resource theories, the hierarchy of contextuality is not a total order — there are GPT systems such that neither is at least as contextual as the other one. It is not even a partial order, because there exist distinct GPT systems which are equivalent. For example, this is the case for all noncontextual systems, such as the classical bit and the classical trit. Therefore, the hierarchy is given by a preorder and it cannot be fully represented by a single numerical value — the "amount of contextuality" — assigned to each GPT system. However, in order to capture certain aspects of the hierarchy it is useful to define quantities which are order-preserving assignments of a number to each resource object. These are called *resource monotones*.

We define a new contextuality monotone that we call *classical excess*. It expresses the minimal error of a univalent simulation of a given GPT system by any classical system. In addition, we show that the optimal success probability for the parity-oblivious-multiplexing (POM) protocol [6] with free classical systems is a monotone.

Previous works on the matter. Resource-theoretic perspective on contextuality has been developed in several works in the past. Most of these focus on the Kochen–Specker notion of contextuality [33]. Despite being related to the generalized notion of contextuality (Kochen–Specker noncontextuality is the conjunction of measurement noncontextuality and outcome determinism for sharp measurements), Kochen–Specker contextuality favors certain frameworks that, instead, are not appropriate for developing a resource theory of generalized noncontextuality. The work of Abramsky, Barbosa, and Mansfield [34] uses a framework whose main objects are empirical models — tables of data, specifying probability distributions over the joint outcomes of sets of compatible measurements. This framework is further developed in [35] and [36] and is based on the sheaf theoretic approach to contextuality introduced in [37]. Existing quantifications of Kochen–Specker contextuality are based on the memory cost [38], the ratio of contextual assignments [39], the relative entropy and contextual cost [40], the contextual robustness [41], the contextual fraction [34], and the rank of contextuality [42]. In [43] a review of several of the previous approaches towards a resource theory of Kochen–Specker contextuality is presented.

The first work on a resource theory of generalized contextuality in prepare-and-measure scenarios was presented by Duarte and Amaral in [44]. They use the generalized-noncontextual polytope characterizing the contextual set of prepared-and-measured statistics defined in [45] to motivate the set of free operations and then define monotones based on known resource quantifiers for contextuality and nonlocality. As an application of such a resource theory, [46] uses it to simplify and robustify proofs of contextuality.

On the use of general probabilistic theories. Unlike Duarte and Amaral, we use the framework of GPTs. A GPT system consists of a collection of states, effects and a probability assigned to each pair of a state and an effect. While noncontextuality is traditionally phrased in terms of operational theories, it can be extended to GPTs [47] and characterized by simplex embeddability [48]. This condition states that a GPT system is noncontextual if and only if it can be embedded within a classical GPT system and thus characterizes the qualitative divide between contextual and noncontextual GPT systems. Our hierarchy of contextuality can be seen as a refinement thereof. For this purpose, we use univalent simulations between GPT systems, which are generalizations of noncontextual ontological models [32].

Structure of the paper. In Section 2, we recall the standard treatment of generalized noncontextuality in the framework of operational theories and ontological models. In order to connect it to the language of GPTs (Section 3.1), we introduce the notion of an operational theory associated to a GPT system in Section 3.2.Readers familiar with ontological models and GPTs may choose to skip directly to Sections 3.3 and 3.4, where we present (univalent) simulations of GPT systems and the excess measure, respectively. We then introduce our hierarchy of contextuality in Section 4.1 and discuss its behaviour for composite systems in Section 4.2. We prove that the classical excess is a resource monotone (Section 4.3) and construct a monotone from the parity-oblivious-multiplexing protocol in Section 4.4. In Section 5.1, we discuss a possible interpretation of the non-free operations as involving information erasure, while in Section 5.2 we discuss the relation to previous works on contextuality and GPTs, namely [32, 44, 48–53]. We conclude with a summary of the results and an outline of possible future research directions in Section 6.

2 Operational theories, ontological models and contextuality

In this work we consider prepare-and-measure scenarios associated with a single system. An *operational theory* associated with a prepare-and-measure scenario is defined by a list of possible preparations, measurements and the probabilities $\mathcal{P}(k|P,M)$ of obtaining the outcome k of the measurement M given that the system is prepared in the preparation P. An *ontological model* of an operational theory provides a realist explanation of the operational predictions of the theory [54]. It does so by stipulating the existence of an *ontic state space* for each given system, denoted with Λ, which is mathematically represented by a (standard Borel) measurable space. Each point $\lambda \in \Lambda$ represents an *ontic state* that describes all the physical properties of the system. The ontological model associates each preparation P in the operational theory with a conditional probability distribution $\mu_P(\lambda) \equiv \mu(\lambda|P)$ over ontic states. We call these *epistemic states* as they represent states of knowledge about the underlying ontic states. Each outcome k given a measurement M, denoted by $[k|M]$, is associated with a conditional probability distribution $\xi(k|\lambda,M)$. The latter corresponds to the probability of obtaining outcome k given that measurement M is implemented on a system in the ontic state λ. An ontological model of an operational theory reproduces the predictions of the theory via the classical law of total probability,

$$\mathcal{P}(k|P,M) = \sum_{\lambda\in\Lambda} \xi(k|\lambda,M)\mu(\lambda|P). \tag{1}$$

An ontological model is preparation noncontextual if operationally equivalent preparation procedures are represented by identical probability distributions in the ontological model [2]. More formally, two preparation procedures P and P' are operationally equivalent if they provide the same operational statistics for all possible measurements, *i.e.*, $\forall M : \mathcal{P}(k|P,M) = \mathcal{P}(k|P',M)$. In this case, we write $P \simeq P'$. An ontological model is preparation noncontextual if any two such preparations are represented by the same epistemic states:

$$P \simeq P' \implies \mu_P = \mu_{P'}. \tag{2}$$

Similarly, two measurement outcomes $[k|M]$ and $[k'|M']$ are operationally equivalent if they give the same statistics for all possible preparations: $\mathcal{P}(k|M,P) = \mathcal{P}(k'|M',P)$ for all preparations

P. In this case, we write $[k|M] \simeq [k'|M']$. An ontological model is measurement noncontextual if any two such measurement outcomes are represented by the same response functions:

$$[k|M] \simeq [k'|M'] \implies \xi_{[k|M]} = \xi_{[k'|M']}. \tag{3}$$

An operational theory is termed *preparation noncontextual* (resp. *measurement noncontextual*) if there exists a preparation noncontextual (resp. measurement noncontextual) ontological model for the theory, while it is termed *preparation contextual* (resp. *measurement contextual*) if it does not admit of a preparation noncontextual (resp. measurement noncontextual) ontological model. It was first proven in [2] that quantum theory is preparation contextual and measurement noncontextual when outcome determinism is not assumed.

We note that the notion of generalized contextuality can be also defined for transformations [2]. In this work we are only concerned with prepare-and-measure scenarios and therefore do not consider transformation contextuality.

3 Contextuality for general probabilistic theories

One can often prove contextuality of a theory by studying statistical behaviours of an individual system A (such as a qubit) in prepare-and-measure scenarios. The information that suffices to describe the possible behaviours constitutes of

- a set Ω of possible states that A can be prepared in, according to the theory in question,
- a set of possible measurements that can be applied to A, each consisting of a collection of effects associated with the measurement outcomes, and
- for every state $\omega \in \Omega$ and every effect e the probability of obtaining e when measurement M is applied to system A prepared in state ω.

This information is commonly expressed as a system in a general probabilistic theory — a GPT system.

In the following we omit any mention of dynamics that the system may undergo. This simplification has no consequence on our discussion of preparation and measurement contextuality.

3.1 GPT systems

Let us now discuss the mathematical description of GPT systems [55–60] and the definition of contextuality in this context.

In the following definition, V_A is a normed vector space and V_A^* is the topological dual of V_A. The canonical pairing function is denoted by $_ \cdot _ : V_\mathsf{A}^* \times V_\mathsf{A} \to \mathbb{R}$, i.e. $e \cdot \omega \coloneqq e(\omega)$.

Definition 1. A **GPT system** A is specified by two non-empty convex sets

$$\Omega_\mathsf{A} \subseteq V_\mathsf{A}, \qquad E_\mathsf{A} \subseteq V_\mathsf{A}^* \tag{4}$$

that are called the **state space** and **effect space**, respectively, while their elements are states and effects of A. We require that E_A contains the **null effect** 0_A and the **unit effect** 1_A satisfying

$$0_\mathsf{A} \cdot v = 0 \quad \forall\, v \in V_\mathsf{A} \qquad \text{and} \qquad 1_\mathsf{A} \cdot \omega = 1 \quad \forall\, \omega \in \Omega_\mathsf{A} \tag{5}$$

respectively. In order for the pairing to have a probabilistic interpretation, we also require $e \cdot \omega \in [0, 1]$ for all effects e and all states ω. Finally, to avoid complications arising from degenerate systems, we require that

$$V_\mathsf{A} = \operatorname{span}(\Omega_\mathsf{A}) \qquad \text{and} \qquad V_\mathsf{A}^* = \operatorname{span}(E_\mathsf{A}) \tag{6}$$

hold, so that states can be distinguished by effects:

$$e \cdot \omega = e \cdot \omega' \quad \forall\, e \in E_A \qquad \implies \qquad \omega = \omega', \tag{7}$$

and similarly that effects can be distinguished by states:

$$e \cdot \omega = e' \cdot \omega \quad \forall\, \omega \in \Omega_A \qquad \implies \qquad e = e'. \tag{8}$$

The set of all GPT systems is denoted by $\mathfrak{G}$ and the subset of finite-dimensional ones by $\mathfrak{G}_{\text{fin}}$.

A GPT system encodes all the information about the statistical behaviour of a physical system in prepare-and-measure scenarios. Let us give a few examples of GPT systems.

Example 2 (Finite-dimensional quantum system as a GPT system). To each Hilbert space $\mathbb{C}^n$ with $n \in \mathbb{N}$, one can associate a quantum GPT system Q_n. The real vector space $V_{\mathsf{Q}_n} \simeq \mathbb{R}^{n^2}$ is the space of Hermitian operators on the underlying Hilbert space. The state space is the set of density operators and the effect space is the set of quantum effects, which are positive semi-definite operators e that satisfy

$$0 \leq e \leq 1, \tag{9}$$

where 1 is the identity operator. The pairing is nothing but the Hilbert–Schmidt inner product:

$$e \cdot \omega = \operatorname{tr}\big(e^\dagger \omega\big). \tag{10}$$

Likewise, every classical probabilistic system is a GPT system [57, §4].

Example 3 (Finite classical GPT systems). We denote the finite n-level classical system (as well as its state space) by Δ_n, with underlying vector space $\mathbb{R}^n$ and state space given by the simplex

$$\Delta_n \coloneqq \operatorname{Conv}\big\{\delta_i \;\big|\; i \in \{1, \ldots, n\}\big\}, \tag{11}$$

where $\{\delta_i\}_i$ is a chosen orthonormal basis of $\mathbb{R}^n$ and Conv denotes the convex hull operation. The effect space consists of all the linear functionals ξ that satisfy

$$\xi \cdot \mu \in [0, 1] \tag{12}$$

for every distribution $\mu \in \Delta_n$. If we think of μ as a column vector, then the possible effects are all the associated row vectors with entries in the unit interval $[0, 1]$. In other words, the effect space is

$$\Delta_n^* \coloneqq \operatorname{Conv}\big\{\rho_\alpha \;\big|\; \alpha \in \{0, 1\}^n\big\}, \tag{13}$$

where

$$\rho_\alpha \coloneqq \sum_{i=1}^{n} \alpha_i \delta_i^* \tag{14}$$

and $\{\delta_i^*\}_i$ is the dual basis to the one above.

Besides finite classical systems, we will also need the notion of a countably infinite one.

Example 4 (Countable classical GPT system). The countable classical system $\Delta_{\mathbb{N}}$ has associated vector space ℓ^1, the Banach space of sequences whose series are absolutely convergent. The state space, also denoted by $\Delta_{\mathbb{N}}$, consists of all probability measures on $\mathbb{N}$. The effect space $\Delta_{\mathbb{N}}^*$ consists of arbitrary sequences in $[0, 1]^{\mathbb{N}}$ and forms a convex subset of ℓ^∞, the space of bounded sequences, which is the dual of ℓ^1.

Besides individual systems, we will occasionally also need to refer to composite ones. In general, the composite of two GPT systems is not unique — see [30, Section 5] for an in-depth discussion of tensor products of GPT systems. However, among all the possible choices, there is a 'minimal composite' of two GPT systems A and B, whose state and effect space merely contain the separable ones. Any meaningful composite of two GPT systems necessarily contains their minimal composite as a subsystem.

Definition 5. Given two GPT systems $\mathsf{A} = (\Omega_\mathsf{A}, E_\mathsf{A}, V_\mathsf{A})$ and $\mathsf{B} = (\Omega_\mathsf{B}, E_\mathsf{B}, V_\mathsf{B})$ we define the **minimal composite** system $\mathsf{A} \otimes \mathsf{B}$ to be given by the separable states and separable effects on A and B. Formally $\mathsf{A} \otimes \mathsf{B} \coloneqq (\Omega_{\mathsf{A}\otimes\mathsf{B}}, E_{\mathsf{A}\otimes\mathsf{B}}, V_{\mathsf{A}\otimes\mathsf{B}})$ where,

$$V_{\mathsf{A}\otimes\mathsf{B}} \coloneqq V_\mathsf{A} \otimes V_\mathsf{B}, \tag{15}$$

$$\Omega_{\mathsf{A}\otimes\mathsf{B}} \coloneqq \operatorname{Conv}\big\{\omega_\mathsf{A} \otimes \omega_\mathsf{B} \;\big|\; \omega_\mathsf{A} \in \Omega_\mathsf{A} \text{ and } \omega_\mathsf{B} \in \Omega_\mathsf{B}\big\}, \tag{16}$$

$$E_{\mathsf{A}\otimes\mathsf{B}} \coloneqq \operatorname{Conv}\big\{e_\mathsf{A} \otimes e_\mathsf{B} \;\big|\; e_\mathsf{A} \in E_\mathsf{A} \text{ and } e_\mathsf{B} \in E_\mathsf{B}\big\}, \tag{17}$$

Of particular interest is the tensor product of an arbitrary GPT system and a classical GPT system, in which case there is a unique choice of a sensible tensor product corresponding to the minimal composite of GPT systems (see Section 4.2 for further discussion about this point).

3.2 Operational theory associated to a GPT system

As we articulate in Section 2, proofs of (preparation) contextuality involve *distinct* preparation procedures, which are nevertheless operationally *equivalent*. Since any two distinct states of a GPT system are operationally *inequivalent*, we cannot identify GPT states with preparation procedures. Instead, we think of a preparation procedure as an ensemble of GPT states, each occurring with a specified probability. Similarly, proofs of measurement contextuality involve *distinct* measurement outcomes, which are operationally *equivalent*. Since any two distinct effects of a GPT system are operationally *inequivalent*, we cannot identify GPT effects with measurement outcomes. Instead, we think of a measurement procedure as a collection of GPT effects, each of which is a possible outcome of the given measurement. See [47, Section 3] for more details.

Such preparations and measurement outcomes then provide a notion of a canonical operational theory associated to a GPT system,[4] to which we can assign ontological models as in Section 2. In this way, we obtain a canonical notion of an ontological model for GPT systems and a corresponding notion of noncontextuality [47, Definitions 2 and 3].

3.3 Simulations of GPT systems

We are, however, interested in comparing GPT systems and not just in distinguishing the noncontextual ones. For this purpose, we use the notion of (univalent) simulations.

Definition 6 ([32, Definition 1])**.** Given two GPT systems A and B, an ϵ-**simulation** of A by B is a pair of

- a multi-valued function $\Gamma : \Omega_{\mathsf{A}} \to \Omega_{\mathsf{B}}$ called **state ϵ-simulation** and
- a multi-valued function $\Theta : E_{\mathsf{A}} \to E_{\mathsf{B}}$ called **effect ϵ-simulation**,

which is empirically adequate up to an error $\epsilon \in [0, \infty)$:

$$\big|e \cdot \omega - f \cdot \gamma\big| \leq \epsilon \qquad \forall\, \omega \in \Omega_{\mathsf{A}},\ \forall\, e \in E_{\mathsf{A}},\ \forall\, \gamma \in \Gamma(\omega),\ \forall\, f \in \Theta(e), \tag{18}$$

preserves convex mixtures:[5]

$$\lambda\Gamma(\omega) + (1-\lambda)\Gamma(\omega') \subseteq \Gamma\big(\lambda\omega + (1-\lambda)\omega'\big) \qquad \text{for all } \lambda \in [0,1] \text{ and } \omega, \omega' \in \Omega_{\mathsf{A}}, \tag{19}$$

and also preserves the null effect, i.e. we have $0_{\mathsf{B}} \in \Theta(0_{\mathsf{A}})$.

A 0-simulation is called an **exact** simulation.[6] If both Γ and Θ are single-valued, we say that the simulation is **univalent**.

One of the immediate consequences of the definition is that the images of states under a state simulation (and of effects under an effect simulation) are non-overlapping. Indeed, by empirical adequacy and the fact that $\Theta(e)$ is always a non-empty subset of E_{B}, we find

$$\begin{aligned} \gamma \in \Gamma(\omega_1) \text{ and } \gamma \in \Gamma(\omega_2) &\implies \forall e \in E_{\mathsf{A}} \ :\ e \cdot \omega_1 = f \cdot \gamma = e \cdot \omega_2 \\ &\implies \omega_1 = \omega_2, \end{aligned} \tag{20}$$

where the second implication is by (7).

Definition 7. Given a pair of GPT systems (A, B), we say that A is **embeddable** within B, denoted by $\mathsf{A} \hookrightarrow \mathsf{B}$, if there exists an exact univalent simulation of type $\mathsf{A} \to \mathsf{B}$. The relation $\hookrightarrow$ is called the **embeddability preorder**.

[4] Conversely, one can always obtain a GPT system from an operational theory. Specifically, each state is an equivalence class of preparation procedures with respect to operational equivalence, and similarly one can define the effects [61].

[5] Analogously, we require that the same holds for mixtures of effects rather than states, just as in [32].

[6] Sometimes, we refer to an exact simulation also just as *a simulation* for brevity.

Indeed, one can easily show that $\hookrightarrow$ is a preorder relation. Specifically, identity maps provide univalent simulations of type $\mathsf{A} \hookrightarrow \mathsf{A}$, which proves the reflexivity of $\hookrightarrow$. To prove that it is a transitive relation, one can use the triangle inequality (see Proposition 12).

It has also been shown that

- a discrete ontological model of a GPT system A provides an exact simulation of A by a finite classical GPT system (and vice versa) [32, Theorem 1],
- a discrete *noncontextual* ontological model of a GPT system A is the same as an exact *univalent* simulation of A by a finite classical GPT system [32, Corollary 1].

In this sense (univalent) simulations generalize (noncontextual) ontological models.

We note that continuous ontological models of GPT systems correspond to simulations by infinite dimensional systems. However, for the purposes of this work and in keeping with the related approaches of [32, 48], we restrict ourselves to discrete ontological models.

Example 8 (Simulations of a bit by a trit). Given a bit Δ_2 and a trit Δ_3 one can define a univalent simulation (Γ, Θ) of Δ_2 by Δ_3 via a direct embedding. Namely, the state and effect simulations act on the bases of $\mathbb{R}^2$ and its dual (see Example 3) via

$$\Gamma\left(\delta_i\right) = \left\{\delta_i\right\}, \qquad \Theta\left(\delta_i^*\right) = \left\{\delta_i^*\right\} \tag{21}$$

and can be extended to all states and effects by linearity. This simulation is univalent since both Γ and Θ are single-valued.

One can also adapt it to a simulation that is not preparation univalent while still being measurement univalent. For instance, we can keep the action on δ_1, δ_1^* and adjust the simulation of the other deterministic state and effect to be

$$\Gamma\left(\delta_2\right) = \left\{\lambda\delta_2 + (1-\lambda)\delta_3 \;\middle|\; \lambda \in [0,1]\right\}, \qquad \Theta\left(\delta_2^*\right) = \left\{\delta_2^* + \delta_3^*\right\} \tag{22}$$

Once we require that Θ also preserves the unit effect, we can extend both to Δ_2 and Δ_2^* by convex-linearity (and using Minkowski sum for the state simulation).

Example 9 (Holevo-Beltrametti-Bugasjki model [61, 62]). Given a GPT system $\mathsf{A} = (\Omega_\mathsf{A}, E_\mathsf{A}, V_\mathsf{A})$ with $n \in \mathbb{N}$ extremal states, the Holevo-Beltrametti-Bugasjki model is given by the following simulation $(\Gamma, \Theta) : \mathsf{A} \to \Delta_n$ where

$$\Gamma\left(\omega_i\right) = \left\{\delta_i\right\}, \qquad \Theta\left(e\right) = \left\{\sum_{i=1}^{n} (e \cdot \omega_i)\delta_i^*\right\} \tag{23}$$

where ω_i is the i-th extremal state of Ω_A and e is an arbitrary effect in E_A. However, it is not a univalent simulation unless Ω_A is a simplex. While the effect simulation is clearly single-valued, to extend the state simulation to all states of Ω_A we need to ensure that property (19) holds. That is, for a given state ω, its simulation $\Gamma(\omega)$ is given by the set of all convex decompositions of ω into extremal states (and applying Γ to each extremal state to land within Δ_n).

3.4 Properties of univalent simulations

The following proposition, which stresses how the composition with a classical system does not affect the univalency of a simulation, is relevant for motivating our notion of contextuality preorder further on (Definition 14) as well as for quantifying contextuality.

Proposition 10. *Let n and k be arbitrary natural numbers. If there is a univalent ϵ-simulation $(\Gamma, \Theta) : \mathsf{A} \otimes \Delta_n \to \mathsf{B}$, then there is also a univalent ϵ-simulation $\mathsf{A} \to \mathsf{B} \otimes \Delta_k$.*

Proof. We show this for $\epsilon = 0$, the general case is analogous. In particular, we define maps $\Gamma_1 : \Omega_\mathsf{A} \to \Omega_\mathsf{B} \otimes \Delta_k$ and $\Theta_1 : E_\mathsf{A} \to E_\mathsf{B} \otimes \Delta_k^*$ via

$$\Gamma_1(\omega) \coloneqq \Gamma(\omega \otimes \delta_1) \otimes \delta_1, \qquad \Theta_1(e) \coloneqq \Theta(e \otimes \delta_1^*) \otimes \delta_1^*. \tag{24}$$

It follows that this is an exact univalent simulation by the assumption that (Γ, Θ) is. Specifically, we have

$$\Theta(e \otimes \delta_j^*) \cdot \Gamma(\omega \otimes \delta_i) = (e \cdot \omega)(\delta_j^* \cdot \delta_i) = \begin{cases} e \cdot \omega & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

which implies $\Theta_1(e) \cdot \Gamma_1(\omega) = e \cdot \omega$. □

By determining the smallest error ϵ for which there is a univalent ϵ-simulation of type $\mathsf{A} \to \mathsf{B}$, we obtain a meaningful notion of how far A is from being embeddable within B.

Definition 11. The **excess** is the function $\varepsilon : \mathfrak{G} \times \mathfrak{G} \to \mathbb{R}$ defined by

$$\varepsilon(A, B) \coloneqq \inf\big\{\epsilon \in [0, \infty) \;\big|\; \text{there is a univalent } \epsilon\text{-simulation } A \to B\big\}, \tag{26}$$

We call $\varepsilon(A, B)$ the A-excess within B.

Note that the value of excess is always in the interval $[0, 1]$. One can show that it cannot exceed 1 by observing that there is always a univalent 1-simulation of type $\mathsf{A} \to \Delta_1$ and a univalent 0-simulation of type $\Delta_1 \to \mathsf{B}$, and using the following proposition.

Proposition 12. *For any three GPT systems A, B, and C, we have the* ***triangle inequality for excess***

$$\varepsilon(A, C) \leq \varepsilon(A, B) + \varepsilon(B, C). \tag{27}$$

Proof. Consider an arbitrary univalent ϵ-simulation $(\Gamma, \Theta) : A \to B$ and an arbitrary univalent γ-simulation $(\alpha, \beta) : B \to C$. We prove the statement by showing that $(\alpha\,\Gamma, \beta\,\Theta)$ is a univalent simulation of type $A \to C$ with error $\epsilon + \gamma$, a result stated as Lemma 3 in [32].

To show that the composite simulation is empirically adequate as expressed by Inequality (18), we can use the triangle inequality. In particular, for all $e_\mathsf{A} \in E_A$ and all $\omega_\mathsf{A} \in \Omega_\mathsf{A}$, we have

$$\begin{aligned} &\big|e_\mathsf{A} \cdot \omega_\mathsf{A} - \beta\,\Theta(e_\mathsf{A}) \cdot \alpha\,\Gamma(\omega_\mathsf{A})\big| \\ &\qquad \leq \big|e_\mathsf{A} \cdot \omega_\mathsf{A} - \Theta(e_\mathsf{A}) \cdot \Gamma(\omega_\mathsf{A})\big| + \big|\Theta(e_\mathsf{A}) \cdot \Gamma(\omega_\mathsf{A}) - \beta\,\Theta(e_\mathsf{A}) \cdot \alpha\,\Gamma(\omega_\mathsf{A})\big| \\ &\qquad \leq \epsilon + \gamma, \end{aligned} \tag{28}$$

where the second inequality follows because $\Theta(e_\mathsf{A})$ is an effect of B and $\Gamma(\omega_\mathsf{A})$ is a state of B, by assumption. □

Note that whenever B is embeddable within C, we have $\varepsilon(\mathsf{B}, \mathsf{C}) = 0$ and the triangle inequality reads

$$\varepsilon(\mathsf{A}, \mathsf{C}) \leq \varepsilon(\mathsf{A}, \mathsf{B}) \tag{29}$$

for any GPT system A. A similar statement follows by setting the first term on the right-hand side of inequality (27) to zero. These two statements can be summarized as the following corollary of Proposition 12.

Corollary 13. *Let A and C be arbitrary GPT systems. With respect to the embeddability preorder among GPT systems, we have the following order-preserving functions:*

$$\varepsilon(\mathsf{A}, _) : (\mathfrak{G}, \hookrightarrow) \to (\mathbb{R}, \geq), \tag{30}$$

$$\varepsilon(_, \mathsf{C}) : (\mathfrak{G}, \hookrightarrow) \to (\mathbb{R}, \leq). \tag{31}$$

In other words, the monotonicity of $\varepsilon(_, \mathsf{C})$ from (31) says that if

$$\varepsilon(\mathsf{A}, \mathsf{C}) > \varepsilon(\mathsf{B}, \mathsf{C}) \tag{32}$$

holds (i.e. the minimal error of simulation by C is strictly higher for A than for B), then there cannot be an exact univalent simulation of type $\mathsf{A} \to \mathsf{B}$. We can thus use the excess within C, for any GPT system C, as a witness of the impossibility to construct an exact univalent simulation.

While simulations are crucial to express a binary division between noncontextual and contextual GPT systems, the embeddability preorder $\hookrightarrow$ is *not* a satisfactory refinement of this binary division. The simplest argument to see this is that all classical GPT systems (which are trivially

noncontextual) should be equivalent in a resource theory of contextuality. However, for any integer m greater than k, the classical GPT system Δ_m *is not* embeddable within Δ_k. Therefore, we cannot think of the embeddability preorder $\hookrightarrow$ as a resource ordering of some resource theory of contextuality. In the following section, we motivate a preorder that further refines the embeddability relation $\hookrightarrow$ and addresses the issue above by stipulating classical GPT systems to be 'free resources'.

4 Hierarchy of contextuality

Recall that our interest in a hierarchy of contextuality of GPT systems stems from the wish to measure how useful contextuality is as a resource in information-theoretic applications. How can one decide whether a given measure of interest is indeed a well-motivated *measure of contextuality*? The resource-theoretic perspective [26, 63], fruitfully applied to studies of entanglement and many other resources in the past, suggests to first motivate a preorder $\succeq$ among the objects of interest. Importantly, some objects may be incomparable according to $\succeq$. In our case, the resource objects are GPT systems and the resource ordering is the advertised hierarchy of contextuality. The interpretation of such an ordering is that the relation $\mathsf{A} \succeq \mathsf{B}$ expresses the statement 'A is at least as contextual as B is'. Given a good motivation for such a hierarchy, one can answer the above question as follows:

> A (real-valued) function on the set $\mathfrak{G}$ of all GPT systems is a measure of contextuality if and only if it is order-preserving with respect to the hierarchy of contextuality $\succeq$.

4.1 Motivation and the resource theory

Let us now list and discuss our desiderata for a sensible hierarchy of contextuality among GPT systems. They are:

1. **Access to classical GPT systems is free.** That is, we have

$$\mathsf{A} \preceq \mathsf{A} \otimes \Delta_n \quad \text{and} \quad \mathsf{A} \otimes \Delta_n \preceq \mathsf{A} \tag{33}$$

   for every GPT system A and every classical GPT system Δ_n. Here, $\otimes$ denotes the minimal composite from Definition 5.

2. **Exact univalent simulations are free.** That is, if A is embeddable within B, then we have $\mathsf{A} \preceq \mathsf{B}$.

Notice that the first condition in Equation (33), i.e. $\mathsf{A} \preceq \mathsf{A} \otimes \Delta_n$, follows from desideratum 2 because every GPT system A is embeddable within $\mathsf{A} \otimes \Delta_n$ (see Proposition 10).

There are several consequences of desiderata 1 and 2 which we want to point the reader's attention to, before we discuss our motivation to introduce these desiderata.

(i) Desideratum 2 implies that the trivial GPT system Δ_1 is *least contextual* in the sense that it is the bottom element of $\preceq$. That is, we have $\Delta_1 \preceq \mathsf{A}$ for every GPT system A.

(ii) Consequently, using the fact that $\Delta_1 \otimes \Delta_n$ is the same GPT system as Δ_n, every classical GPT system is also at the bottom by Desideratum 1.

(iii) Finally, using the fact that noncontextual systems are simplex embeddable and Desideratum 2, we conclude that all noncontextual GPT systems are equivalent to each other and lie at the bottom of the hierarchy.

Motivation for Desideratum 1. Contextuality, as we are thinking of it, is a property that expresses the degree to which a system's behaviour escapes a purely classical explanation. In this sense, considering additional classical systems should not affect this property. An important point to note is that within the tensor product $\mathsf{A} \otimes \Delta_n$ (Definition 5), only correlations involving separable states between the generic GPT system A and the classical system are allowed. We can

think of behaviours of $\mathsf{A} \otimes \Delta_n$ as dilations of those of A [64], given a classical environment variable Δ_n.

If there were classical environmental variables, whose knowledge reduces the contextuality present in A, then the contextuality of A would not be an authentic feature of A, rather a consequence of one's too narrow focus on A as opposed to its dilation $\mathsf{A} \otimes \Delta_n$. We can also support this argument on the technical side. Indeed, by Proposition 10, any noncontextual model of $\mathsf{A} \otimes \Delta_n$ provides also a noncontextual model of A. In other words, A is noncontextual if and only if $\mathsf{A} \otimes \Delta_n$ is.[7] The hierarchy of contextuality ought to capture this fact.

Motivation for Desideratum 2. We can also think of contextuality (of a GPT system) as the inability to provide a noncontextual model (i.e. a univalent simulation by a classical system) for its statistical behaviours. If a GPT system A can be exactly simulated by a GPT system B via a univalent simulation, then any obstruction to constructing such a noncontextual model for A must already be present in B. In this sense, we think of B as being "at least as contextual" as A is. Indeed, provided with a noncontextual model of B, i.e. an exact univalent simulation of B by a classical GPT system, one can construct a noncontextual model of A by composition with the simulation of A by B.

The preorder relation among finite-dimensional GPT systems in $\mathfrak{G}_{\text{fin}}$ that satisfies precisely these requirements is the following one.

Definition 14. We say that B is at least as contextual as A, denoted $\mathsf{A} \preceq \mathsf{B}$, whenever there is an exact univalent simulation of A by $\mathsf{B} \otimes \Delta_n$ for some finite-dimensional classical GPT system Δ_n. The preordered set $(\mathfrak{G}_{\text{fin}}, \succeq)$ is called the **hierarchy of contextuality**.

In other words, we define the hierarchy via embeddability given an additional classical system:

$$\mathsf{A} \preceq \mathsf{B} \quad \Longleftrightarrow \quad \exists\, n \in \mathbb{N} \text{ such that } \mathsf{A} \hookrightarrow \mathsf{B} \otimes \Delta_n \text{ holds.} \tag{34}$$

To show that the hierarchy $\preceq$ is a preorder relation, we can use a resource-theoretic perspective, i.e. we can show that the set of free operations in the following resource theory is closed under composition.

Definition 15. The **resource theory of GPT-contextuality** is defined as follows. Its objects are finite-dimensional GPT systems in $\mathfrak{G}_{\text{fin}}$. The set of transformation from B to A is identified with exact simulations of type $\mathsf{A} \to \mathsf{B} \otimes \Delta_n$. Two such simulations

$$(\Gamma, \Theta) : \mathsf{A} \to \mathsf{B} \otimes \Delta_m \quad \text{and} \quad (\alpha, \beta) : \mathsf{B} \to \mathsf{C} \otimes \Delta_k \tag{35}$$

can be composed (sequentially) to produce a simulation

$$(\alpha\,\Gamma, \beta\,\Theta) : \mathsf{A} \to \mathsf{C} \otimes \Delta_{m \cdot k} \tag{36}$$

via the canonical isomorphism between $\Delta_m \otimes \Delta_k$ and $\Delta_{m \cdot k}$. Free transformations are those simulations that are also *univalent*.

This gives rise to a concrete resource theory in the sense of [26, Definition 3.10]. Moreover, the resulting resource ordering coincides with our hierarchy of contextuality. In a general resource theory, the existence of a free transformation from a resource object r to a resource object s defines the resource ordering denoted by $s \preceq r$, i.e. r is at least as good as a resource than s is.

In our work, a simulation of type $\mathsf{C} \to \mathsf{D}$ is interpreted as providing effective access to system C to an agent with access to D. In this sense, access to D grants the agent at least as many capabilities as access to C. For this reason, a simulation of type $\mathsf{A} \to \mathsf{B} \otimes \Delta_n$ is identified with a resource-theoretic transformation from B to A and not vice versa. Note, however, that it is not in general realisable by a physical transformation from B to A in the GPT sense. A physical transformation $\mathsf{B} \to \mathsf{A}$ is given by a linear map $M : V_{\mathsf{B}} \to V_{\mathsf{A}}$ such that $M(\Omega_{\mathsf{B}}) \subseteq \Omega_{\mathsf{A}}$ and $M^*(E_{\mathsf{A}}) \subseteq E_{\mathsf{B}}$. For further discussion see Sections A and 5.1.

[7] Note that under the minimal tensor product of Definition 5 allowing access to arbitrary noncontextual systems (and not just classical systems) would lead to the same hierarchy, which follows from the fact that access to classical systems already places all noncontextual systems at the bottom of the hierarchy.

4.2 Contextuality of composite systems

When studying a resource theory, it is often also important to specify how resource objects are composed to obtain a joint resource. As discussed in Section 3.1, GPT systems do not have a canonical tensor product. However, any sensible choice of a tensor product $\tilde{\otimes}$ should give rise to a symmetric monoidal category with exact simulations as morphisms [30, Section 5.2]. Since the transformations in our resource theory of GPT-contextuality are not merely simulations of GPT systems, but also involve additional classical systems, it takes a bit more work to show that they (and the free transformations) are closed under parallel composition. We prove this in the rest of this section for the case of free operations.

Consider two exact univalent simulations of types $\mathsf{A}_1 \to \mathsf{B}_1 \otimes \Delta_m$ and $\mathsf{A}_2 \to \mathsf{B}_2 \otimes \Delta_k$ respectively. The assumption that $\tilde{\otimes}$ is the monoidal product of a symmetric monoidal category then means that there is also an exact univalent simulation of type

$$\mathsf{A}_1 \,\tilde{\otimes}\, \mathsf{A}_2 \to (\mathsf{B}_1 \otimes \Delta_m) \,\tilde{\otimes}\, (\mathsf{B}_2 \otimes \Delta_k)\,. \tag{37}$$

Assuming that there is a unique composite of any GPT system with a classical one given by the canonical composite of Definition 5 (see [65, 66] for a derivation of the uniqueness of the composition with a classical system under the assumption of local tomography and [67] for an alternative composition of classical systems without this assumption) and that $\tilde{\otimes}$ is associative, the codomain can be expressed as

$$\begin{aligned}(\mathsf{B}_1 \otimes \Delta_m) \,\tilde{\otimes}\, (\mathsf{B}_2 \otimes \Delta_k) &= (\mathsf{B}_1 \,\tilde{\otimes}\, \Delta_m) \,\tilde{\otimes}\, (\mathsf{B}_2 \,\tilde{\otimes}\, \Delta_k)\\ &\cong (\mathsf{B}_1 \,\tilde{\otimes}\, \mathsf{B}_2) \,\tilde{\otimes}\, (\Delta_m \,\tilde{\otimes}\, \Delta_k)\\ &\cong (\mathsf{B}_1 \,\tilde{\otimes}\, \mathsf{B}_2) \,\tilde{\otimes}\, \Delta_{m\cdot k}\\ &= (\mathsf{B}_1 \,\tilde{\otimes}\, \mathsf{B}_2) \otimes \Delta_{m\cdot k},\end{aligned} \tag{38}$$

where the third equality uses the standard isomorphism between $\Delta_m \otimes \Delta_k$ and $\Delta_{m\cdot k}$. In the first and final steps we also use the assumption that composition with a classical system is unique. Consequently, we get a parallel composite of the two free transformations as an exact univalent simulation of type

$$\mathsf{A}_1 \,\tilde{\otimes}\, \mathsf{A}_2 \to (\mathsf{B}_1 \,\tilde{\otimes}\, \mathsf{B}_2) \otimes \Delta_{m\cdot k}. \tag{39}$$

As a consequence, the parallel composition of resources (i.e. GPT systems) via $\tilde{\otimes}$ respects the hierarchy of contextuality in the sense that we have

$$\mathsf{A}_1 \preceq \mathsf{B}_1 \;\wedge\; \mathsf{A}_2 \preceq \mathsf{B}_2 \quad\Longrightarrow\quad \mathsf{A}_1 \,\tilde{\otimes}\, \mathsf{A}_2 \preceq \mathsf{B}_1 \,\tilde{\otimes}\, \mathsf{B}_2, \tag{40}$$

for arbitrary GPT systems A_i and B_i. Condition (40) is part of the definition of a resource theory via ordered commutative monoids in [68]. It means that we can consistently apply the hierarchy of contextuality also to express the contextuality of composite GPT systems, as long as their composition satisfies a few basic properties as described above.

4.3 Quantifying contextuality via the classical excess

As is common in resource theories [69], we can use monotones — i.e. order-preserving functions from resources to numbers — to study the hierarchy of contextuality. On the one hand, these can provide a lens through which to extract properties of the preorder. On the other hand, they can be used as quantitative measures of contextuality. The latter perspective is particularly useful when the monotones come equipped with an (operational) interpretation that allows one to identify *which aspect of contextuality* they are measuring.

In order to construct a specific contextuality monotone, we will make use of the excess function from Definition 11. Of particular interest is $\varepsilon(\mathsf{A}, \Delta_{\mathbb{N}})$, the A-excess within the countable classical GPT system (Example 4). For brevity, we call it the **classical excess** of A. A crucial ingredient in the following proof is that $\Delta_{\mathbb{N}}$ is isomorphic to $\Delta_{\mathbb{N}} \otimes \Delta_n$ for any finite n.

Lemma 16. *For any GPT system* A *and any finite-dimensional classical GPT system* Δ_n*, we have*

$$\varepsilon(\mathsf{A} \otimes \Delta_n, \Delta_{\mathbb{N}}) = \varepsilon(\mathsf{A}, \Delta_{\mathbb{N}}) \tag{41}$$

Proof. Firstly, the inequality

$$\varepsilon(\mathsf{A} \otimes \Delta_n, \Delta_{\mathbb{N}}) \geq \varepsilon(\mathsf{A}, \Delta_{\mathbb{N}}) \tag{42}$$

is a direct consequence of Proposition 10 and the definition of excess.

For the converse inequality, note that any univalent ϵ-simulation $(\Gamma, \Theta) : \mathsf{A} \to \Delta_{\mathbb{N}}$ can be canonically extended to a univalent ϵ-simulation

$$\big(\Gamma \otimes \mathrm{id}, \Theta \otimes \mathrm{id}\big) \ : \ \mathsf{A} \otimes \Delta_n \to \Delta_{\mathbb{N}} \otimes \Delta_n, \tag{43}$$

so that we get

$$\varepsilon(\mathsf{A}, \Delta_{\mathbb{N}}) \geq \varepsilon(\mathsf{A} \otimes \Delta_n, \Delta_{\mathbb{N}} \otimes \Delta_n). \tag{44}$$

Furthermore, there is (an exact) univalent simulation $\Delta_{\mathbb{N}} \otimes \Delta_n \to \Delta_{\mathbb{N}}$ given on states by the linear map

$$\begin{aligned} \ell^1 \otimes \mathbb{R}^n &\to \ell^1 \\ (\omega_i)_{i\in\mathbb{N}} \otimes \delta_j &\mapsto \big(\omega_{i/j}\delta_{j|i}\big)_{i\in\mathbb{N}}, \end{aligned} \tag{45}$$

where $(\omega_i)_{i\in\mathbb{N}}$ is an element of ℓ^1, δ_j is a basis vector in $\mathbb{R}^n$, and $\delta_{j|i}$ returns 1 if j divides i and 0 otherwise. Using basically the same linear map for effects, merely extended to $\ell^\infty \otimes \mathbb{R}^n \to \ell^\infty$, gives the required simulation. Consequently, we have

$$\varepsilon(\Delta_{\mathbb{N}} \otimes \Delta_n, \Delta_{\mathbb{N}}) = 0, \tag{46}$$

so that, by Inequality (44) and Proposition 12, we get

$$\begin{aligned} \varepsilon(\mathsf{A}, \Delta_{\mathbb{N}}) &\geq \varepsilon(\mathsf{A} \otimes \Delta_n, \Delta_{\mathbb{N}} \otimes \Delta_n) + \varepsilon(\Delta_{\mathbb{N}} \otimes \Delta_n, \Delta_{\mathbb{N}}) \\ &\geq \varepsilon(\mathsf{A} \otimes \Delta_n, \Delta_{\mathbb{N}}) \end{aligned} \tag{47}$$

as required. □

Theorem 17. *The function $\varepsilon(_, \Delta_{\mathbb{N}}) : \mathfrak{G} \to \mathbb{R}$, which maps a GPT system to its excess within $\Delta_{\mathbb{N}}$, is order-preserving. That is, we have*

$$\mathsf{A} \preceq \mathsf{B} \quad \Longrightarrow \quad \varepsilon(\mathsf{A}, \Delta_{\mathbb{N}}) \leq \varepsilon(\mathsf{B}, \Delta_{\mathbb{N}}). \tag{48}$$

Proof. By Definition 14, Corollary 13 (specialized to $\mathsf{C} = \Delta_{\mathbb{N}}$), and Lemma 16 respectively, we have

$$\begin{aligned} \mathsf{A} \preceq \mathsf{B} \quad &\Longleftrightarrow \quad \exists\, n \ \text{ such that } \ \mathsf{A} \hookrightarrow \mathsf{B} \otimes \Delta_n, \\ &\Longrightarrow \quad \exists\, n \ \text{ such that } \ \varepsilon(\mathsf{A}, \Delta_{\mathbb{N}}) \leq \varepsilon(\mathsf{B} \otimes \Delta_n, \Delta_{\mathbb{N}}) \\ &\Longleftrightarrow \quad \varepsilon(\mathsf{A}, \Delta_{\mathbb{N}}) \leq \varepsilon(\mathsf{B}, \Delta_{\mathbb{N}}). \end{aligned} \tag{49}$$

This gives us the statement of the theorem. □

Remark 18 (Excess within a finite simplex). An important feature of the monotone $\varepsilon(\mathsf{A}, \Delta_{\mathbb{N}})$ is that it can sometimes be computed by considering the A-excess within Δ_k for a finite k. This happens whenever its value $\varepsilon(\mathsf{A}, \Delta_m)$ remains constant for all classical systems with m larger than k. For instance, the noncontextual GPT system corresponding to the toy bit in Spekkens toy theory [70] is embedabble within Δ_4 with vanishing error, and hence for any larger m its excess within Δ_m has to remain 0.

4.4 Parity oblivious multiplexing success probability with free classical resources as a measure of contextuality

Having introduced the hierarchy of contextuality as well as a new contextuality measure in the form of the classical excess, we now build a measure of generalized contextuality based on the parity oblivious multiplexing (POM) [10] game. The latter is a protocol that is powered by preparation contextuality and has raised significant attention in recent years [71–78].

In the ***n*-bit Parity Oblivious Multiplexing (POM) game**, Alice is given an n-bit string, whose possible values she encodes in a GPT system A. In particular she chooses a state ω_x for possible string value x. The system is then transmitted to Bob is additionally given an integer

$y \in \{1, 2, \ldots, n\}$. Bob's task is to guess the value x_y of the y^{th} bit of x by performing a two-outcome measurement M_y. His guess is denoted by $b \in \{0, 1\}$.

The task has a supplementary constraint called parity obliviousness (PO). Namely, Alice cannot communicate the parity of the string, denoted by $p(x)$, to Bob. The PO condition can be phrased in terms of Alice's chosen states as

$$\sum_{x|p(x)=0} \omega_x = \sum_{x|p(x)=1} \omega_x. \tag{50}$$

Note that the string x is assumed to be uniformly distributed, so that the left-hand side of Equation (50) can be interpreted (up to a factor of 2^{n-1}) as the average state sent by Alice given that the characters of x add up to 0 modulo 2 and similarly for the right-hand side.

The average success probability is given by

$$\sum_{x,y} \frac{1}{2^n n} p(b = x_y \,|\, \omega_x, M_y), \tag{51}$$

where 2^{-n} is the probability of Alice receiving string x, $1/n$ is the probability of Bob receiving y, and $p(b = x_y \,|\, \omega_x, M_y)$ is the probability that Bob's guess is correct in a given run.

It is shown in [10, Theorem 2] that if a system A is noncontextual, then the **optimal** (i.e. maximal) **success probability** for the n-bit POM, denoted by $\mathfrak{p}_n$, is upper bounded by

$$\mathfrak{p}_n(\mathsf{A}) \leq \frac{n+1}{2n} = \mathfrak{p}_n(\Delta_\infty). \tag{52}$$

The success probability for POM can be thus used as a witness of contextuality.

We now prove that for the embeddability preorder the POM success probability is also order-preserving.

Lemma 19. *Let n be any integer. If A is embeddable within B, then we have $\mathfrak{p}_n(\mathsf{A}) \leq \mathfrak{p}_n(\mathsf{B})$.*

Proof. Let us consider any choice of states ω_x (for $x \in \{0,1\}^n$) which obey the parity obliviousness constraint (50). Since we have $\mathsf{A} \hookrightarrow \mathsf{B}$, for any choice of effects e_y and states ω_x on A obeying the parity obliviousness constraint, the states $\Gamma(\omega_x)$ on B also obey it by the linearity of the single-valued state simulation Γ. Moreover, the empirical adequacy (Equation (18)) of the exact simulation implies $e_y \cdot \omega_x = \Gamma(e_y) \cdot \Gamma(\omega_x)$. Thus, any strategy on A can be implemented on B with equal success probability. This implies $\mathfrak{p}_n(\mathsf{A}) \leq \mathfrak{p}_n(\mathsf{B})$. □

However, it is not a resource monotone in our resource theory of GPT-contextuality.

Lemma 20. *The optimal success probability for POM is not a monotone with respect to the hierarchy of contextuality $\preceq$.*

Proof. We define a *noisy bit* to be any system with the same states as the classical bit Δ_2, but whose effects form a proper subset of Δ_2^*. In particular, a noisy bit has no measurement that distinguishes the two extremal states δ_1 and δ_2 with certainty. Noisy bits can be parametrized by a noise parameter $\alpha \in (0, 1/2]$, giving rise to a GPT system Δ_2^α whose extremal effects χ_1 and χ_2 satisfy

$$\chi_i \cdot \delta_j = \begin{cases} 1-\alpha & \text{if } i = j \\ \alpha & \text{if } i \neq j. \end{cases} \tag{53}$$

By the desiderata 1 and 2, a noisy bit Δ_2^α and a bit Δ_2 must satisfy $\Delta_2^\alpha \preceq \Delta_2$ and $\Delta_2 \preceq \Delta_2^\alpha$ since they are both noncontextual. The optimal strategy for POM given a (noisy) bit consists of encoding the first bit x_1 of the string x in its extremal states δ_1 and δ_2 [10, Lemma 1]. With probability $1/n$ Bob is asked to guess this specific bit, which he can do successfully with probability $(1-\alpha)$. With probability $(n-1)/n$ Bob is asked to guess any of the remaining bits from x, for which he can do no better than a uniformly random guess, i.e. the success probability is $1/2$. Hence, the optimal success probability is

$$\mathfrak{p}_n(\Delta_2^\alpha) = \frac{1-\alpha}{n} + \frac{n-1}{2n} = \frac{n+1-2\alpha}{2n}, \tag{54}$$

which is strictly smaller than $\mathfrak{p}_n(\Delta_2) = (n+1)/2n$ whenever the noise parameter α is greater than 0. This means that $\Delta_2 \preceq \Delta_2^\alpha$ does not imply $\mathfrak{p}_n(\Delta_2^\alpha) \geq \mathfrak{p}_n(\Delta_2)$. □

We can see that the POM success probability cannot serve as a monotone because the POM game does not allow the use of classical systems for free. This can be remedied by the so-called generalized yield construction [69, Section 3.1].

Definition 21. Given a GPT system A we define the optimal **POM success probability with free classical resources** as:

$$\mathfrak{p}_n\text{-yield}(\mathsf{A}) \coloneqq \sup_{d\in\mathbb{N}} \mathfrak{p}_n(\mathsf{A}\otimes\Delta_d). \tag{55}$$

Proposition 22. $\mathfrak{p}_n$-yield *is a monotone on the resource theory of GPT-contextuality.*

Proof. Let us assume $\mathsf{A}\preceq\mathsf{B}$, which implies that there is an $m\in\mathbb{N}$ such that there is a univalent simulation $\mathsf{A}\otimes\Delta_d\to\mathsf{B}\otimes\Delta_{d\cdot m}$ for every $d\in\mathbb{N}$. By Lemma 19, we obtain $\mathfrak{p}_n(\mathsf{A}\otimes\Delta_d)\leq\mathfrak{p}_n(\mathsf{B}\otimes\Delta_{d\cdot m})$ for every $d\in\mathbb{N}$, which yields the inequality below:

$$\sup_{d\in\mathbb{N}} \mathfrak{p}_n(\mathsf{A}\otimes\Delta_d) \leq \sup_{d\in\mathbb{N}} \mathfrak{p}_n(\mathsf{B}\otimes\Delta_{d\cdot m}). \tag{56}$$

Moreover,

$$\sup_{d\in\mathbb{N}} \mathfrak{p}_n(\mathsf{B}\otimes\Delta_{d\cdot m}) = \sup_{d\in\mathbb{N}} \mathfrak{p}_n(\mathsf{B}\otimes\Delta_d), \tag{57}$$

which implies

$$\mathfrak{p}_n\text{-yield}(\mathsf{A}) \leq \mathfrak{p}_n\text{-yield}(\mathsf{B}), \tag{58}$$

using Definition 21. □

This proof is an instance of the proof of the generalized yield construction in [26, Theorem 4.21], so that $\mathfrak{p}_n$-yield can be meaningfully interpreted as a (generalized) yield monotone.

5 Discussion

This section is divided into two subsections. In the first, we discuss whether our resource-theoretic perspective on contextuality could offer an avenue to understand contextuality as an expression of information erasure. In the second, we discuss how our work relates to the main studies on which it builds, as well as to the resource theory of generalized contextuality developed in [44].

5.1 Contextuality and information erasure

The resource-theoretic perspective on contextuality [10, 12, 14–18, 44, 47, 74, 79] highlights the following aspect: the existence of a noncontextual model (i.e. noncontextuality) of a physical system imposes constraints on the type of information that this system can carry. This is in essence why one can show that its violation (i.e. contextuality) can act as a resource for information-theoretic tasks. Our resource theory of GPT-contextuality aims to give a more precise description of the nature and properties of this information-theoretic manifestation of contextuality as compared to previous works that often focus merely on the distinction between contextual and noncontextual systems.

From a more foundational perspective, a contextual ontological model gives a classical description of an operational theory that postulates ontological distinctions that are in principle inaccessible. There are two types of responses to the phenomenon of contextuality (e.g. that of quantum systems). One is to reject ontological models as the appropriate framework to explain the operational features of the theory. The other is to accept the possibility that a contextual ontological model provides a feasible realist explanation of the empirical predictions of the theory.

The latter, however, begs for an explanation of the fine-tuning associated to it [4]. How can we understand that ontological distinctions disappear at the operational level? One could explain the presence of fine-tuning implied by contextuality if there was a physical process responsible for this disappearance. It would map the ontological theory to the operational theory, effectively erasing the extra information that is supposed to be inaccessible at the level of the operational theory. We

refer to this hypothetical process, associated to any ontological model, as the *revelation map*, for short.

One could then ask: can revelation maps explain the detailed information-theoretic properties of contextuality? To tackle this question, we may take our resource theory of GPT-contextuality as an expression of the information-theoretic properties of contextuality. In this sense, the question then becomes: can the resource theory of GPT-contextuality be alternatively interpreted as a resource theory of erasure for revelation maps? We cannot currently answer this question, since a resource theory of erasure for revelation maps has not yet been developed. Nevertheless, we believe the question is worthwhile and we outline a few relevant considerations below.

Explaining fine-tunings as emergent from yet undiscovered physical mechanisms. The presence of fine-tunings in a contextual ontological model gives a conspiratorial connotation to the realist explanation of the theory it provides. One attempt to justify the presence of such fine-tunings, by searching for new physical mechanisms, dates back to Valentini's variant of Bohmian mechanics [80]. There, he introduces a notion of quantum equilibrium as the reason why superluminal signaling does not manifest in quantum theory, despite the nonlocality of its underlying ontological model. This picture predicts that outside of the quantum equilibrium, it is possible to observe faster than light signaling. Therefore, the fine-tuned nature of no-signaling in Bohmian mechanics is explained just as an emergent feature of the quantum equilibrium and it is not universally valid. We cannot avoid noticing how radical such explanations of fine-tunings rooted in undiscovered physical mechanisms are. They imply that an established physical principle, such as the principle of no-signaling, is violated at the fundamental level. In the case of contextuality, the physical mechanism explaining the emergence of the operational equivalences would entail the existence of measurements that can distinguish behaviours that are deemed indistinguishable by quantum theory.

Explaining contextuality through information erasure. In Valentini's work, the quantum equilibration process is responsible for the emergence of no-signalling — the fine-tuned feature associated with nonlocality. What hypothetical physical mechanism could be responsible for the emergence of operational equivalences — the fine-tuned feature associated with contextuality?

It would have to be a process that involves a kind of information erasure. The information erased is the information about distinctions at the ontological (e.g. fundamental) level, which cannot be stored in systems of the operational (e.g. effective) theory that lacks these distinctions. By Landauer's principle, we can then associate an increase in entropy between the fundamental and effective levels. Such a process of information erasure would not only provide an explanation for the problematic fine-tuning associated with contextuality but would also be associated to a potentially detectable heat dissipation. This heat would signify that, indeed, there are distinctions at the fundamental level which are not present at the effective level. One could even hypothesise that the information erasure is a physical process occurring over time. That is, during the preparation of a quantum system there may be a timescale before which the system is described by the fundamental (and noncontextual) theory. At longer timescales, once the erasure has occurred, the system can only be described by the effective (contextual) theory.

Empirical nature of the fundamental theory. In the above hypothetical account of Bohmian mechanics, the theory of Bohmian particles is seen as a more fundamental one than quantum theory.

In order for a similar explanation of the fine-tuning associated with contextuality to be testable, the more fundamental theory must be in principle accessible. Only then would we expect entropy increase and heat transfer as a result of the erasure in the revelation map. This is in accordance with the view presented by Müller and Garner in [32], but it is not in opposition to the standard use of ontological models to study contextuality [2].

Indeed, if one tries to find evidence for the physical erasure process we hypothesize here and fails, this gives more credence to the common assumption that any ontological distinctions, which are not present in the operational theory, are in principle indistinguishable. On the other hand, if such evidence is found, the detailed study of contextuality in quantum theory provides invaluable clues towards finding the more fundamental theory, from which quantum theory emerges.

Ontological models as physical erasure processes. In case the fundamental theory in question is classical, one is justified in using standard ontological models to describe how the effective theory is simulated by the fundamental one. Namely, an ontological model of (the operational theory associated to) a GPT system A (Section 3.2) is equivalent to a simulation of A by a classical system Δ_n. The classical system Δ_n is interpreted as the fundamental system and A as the effective description thereof. A simulation $\mathsf{A} \to \Delta_n$ specifies all behaviours of the fundamental theory Δ_n that are compatible with a given behaviour of the effective theory A.

Whether we can interpret A as a description of Δ_n that emerges as a result of a revelation map from Δ_n to A depends on the kind of revelation maps that we consider. For example, if it supposed to be a pair of partial functions (one for states, one for effects) $f : \Delta_n \to \mathsf{A}$, then the associated simulation would be given by preimages under f. We could obtain an arbitrary simulation in this way.

However, many simulations cannot be obtained like this if we further require the revelation map f to be a *physical GPT transformation* (Definition 23). Namely, we say that a simulation (Γ, Θ) of A by Δ_n admits of a *physical realisation* if there exists a linear map $M : \Delta_n \to \mathsf{A}$, such that $\Gamma(\omega) \subseteq M^{-1}(\omega)$ for all $\omega \in \Omega_{\mathsf{A}}$ and $M^*(e) \in \Theta(e)$ for all $e \in E_{\mathsf{A}}$. See Section A for more details.

For instance, while there is a noncontextual ontological model of the Spekkens toy bit [70] by two classical bits, i.e. a univalent simulation by Δ_4, there cannot exist a physical realisation of this simulation, as we prove in Section A.

Nevertheless, there are ontological models with physical realisations. For any GPT system A with a finite set $\{\omega_i\}_{i=1}^n$ of convexly extremal states, we can define the Holevo-Beltrametti-Bugajski (HBB) model of A (Example 9), which may be contextual. In Lemma 27, we show that every HBB model, viewed as a simulation of A by a classical GPT system Δ_n, admits of a physical realisation $M : \Delta_n \to \mathsf{A}$. At the level of states, M maps a given extremal state of the simplex Δ_n to the corresponding extremal state ω_i of Ω_{A}.

Erasure associated with ontological models. By analyzing physical GPT transformations like M above, we can understand why information erasure in the context of revelation maps cannot be the same as the erasure in the context of thermodynamics. While the latter can be understood as a coarse-graining operation, physical maps associated to contextual ontological models are not necessarily of this kind.

To see this, consider a contextual ontological model of a GPT system A, which is given by preimages under a physical GPT transformation $f : \Delta_d \to \mathsf{A}$. Since A is contextual, there is no coarse-graining (i.e. idempotent) process on Δ_d whose image is A [32, Lemma 11]. Hence, the physical GPT transformation f cannot be understood simply as a coarse-graining operation, such as the erasure that occurs between fundamental and effective levels in statistical mechanics.

A more concrete perspective is offered by the HBB map M which explicitly models the preparation of the system A via the classical control Δ_n. The classical degree of freedom is used to select the extremal state of A to be prepared. Thus, while M can be used to model a physical process in the laboratory (e.g. the preparation of A conditioned on a classical variable), it does not correspond to a simple coarse-graining process applied to the classical degree of freedom.

5.2 Relation with previous works on contextuality and GPTs

In the following section we relate the resource theory of generalized contextuality introduced in this work to previous works on generalized contextuality in the frameworks of GPTs and contrast it to the existing proposed resource theory of contextuality of [44].

On the framework of Müller and Garner. On the technical level, our work makes use of the tools developed in [32]. For instance, the concept of univalent simulations of GPT systems plays an especially important role. Even though we use similar tools, our goal is quite different. Müller and Garner use the concept of univalent simulations to formulate a notion of generalized noncontextuality that applies to generic effective and corresponding fundamental theories. The effective theory is the theory emerging from the operational statistics of the experiment and the

fundamental theory is a fine-grained theory of the effective ones (so, not just the simplicial theory, as in the case of standard noncontextuality).

They use such noncontextuality as a plausibility criterion for testing the validity of the fundamental theory given the effective theory. This approach becomes particularly meaningful when the fundamental theory is assumed to be quantum theory, since it allows the authors to provide experimental tests of quantum theory. Related works which employ an alternative definition of noncontextuality are the ones of Gitton and Woods [49] and [50]. For ongoing debate on the merits of the different approaches see [50] and [81].

On simplex embeddability. Other influential sources of inspiration for our work are [48, 52, 53]. As we already mentioned in the introduction, [48] introduces the notion of simplex embeddability as the geometrical criterion to assess whether a GPT system is noncontextual.[8][9] The criterion is extended to accessible GPT fragments in [52], where the latter correspond to more general mathematical objects than GPTs and characterize generic prepare-and-measure experimental setups. The work of [53] provides an algorithm for testing contextuality in any prepare-and-measure scenario. In particular, if it exists, it returns an explicit noncontextual model for the scenario and, if not, it provides the minimum amount of (depolarizing) noise which would be required until a noncontextual model would become possible. They call this measure of contextuality the *robustness of nonclassicality*. The latter is related to our notion of error of univalent simulation, but it is different insofar as it requires a specific noise channel (e.g. depolarizing noise channel), whereas the error of univalent simulation does not require that. Moreover, the robustness of nonclassicality is not defined as a measure within a well defined resource theory.

On accessible GPT fragments. It has been argued in [52] that in realistic experiments GPT systems are generally not appropriate to study contextuality. In particular, the following discrepancy arises: A GPT system obtained from an operational description of the experiment provides states and effects observable *in practice* in this experiment, but does not limit the *in principle* implementable ones. However, the preparation equivalences preserved by a preparation-noncontextual ontological model (see Section 2) are *in principle* operational equivalences.

When studying preparation contextuality only, this discrepancy can be dealt with. Namely, one can first set-up the effect space E_{A} in a non-operational way, as a hypothesized set of in principle possible effects. Importantly, one has to identify each actual measurement outcome arising in the experiment with a (distribution over) the elements of E_{A}. Their means specify a subset $\tilde{E}_{\mathsf{A}} \subseteq E_{\mathsf{A}}$ of the so-called *accessible* effects, which however *play no role* in the study of preparation contextuality. The state space Ω_{A} is the set of equivalence classes of preparation procedures, but only with respect to in principle operational equivalence. That is, states that are indistinguishable by accessible effects, but could be distinguished by a hypothesized effect in $E_{\mathsf{A}} \setminus \tilde{E}_{\mathsf{A}}$ ought not be identified in Ω_{A}. With this set-up, one can study preparation contextuality as we discuss shortly in Section 3.3 in terms of the GPT system $(\Omega_{\mathsf{A}}, E_{\mathsf{A}})$.

Such a treatment is not possible when studying both preparation and measurement contextuality simultaneously. One has to replace the notion of a GPT system by the notion of an *accessible fragment* of a GPT system introduced in [52]. Simply put, this is a pair of a GPT system $(\Omega_{\mathsf{A}}, E_{\mathsf{A}})$ together with a subsystem $(\tilde{\Omega}_{\mathsf{A}}, \tilde{E}_{\mathsf{A}})$ thereof, which importantly can violate the non-degeneracy condition (7), and therefore are not necessarily GPT system. The enveloping system describes states and effects that can arise in principle, while the subsystem describe those that are achievable in the particular experiment, whose contextuality we are interested in.

[8] As a note our notion of the canonical ontological model associated to a GPT system differs from the notion of ontological model of a GPT system of [48] since we define the preparations to be the set of ensembles of states, whereas they define the preparations to be the set of states (hence every ontological model of a GPT system is noncontextual in their work).

[9] Another work that is closely related to [48] is that of Shahandeh [51]. However, the criterion of classicality introduced there additionally requires the simplex to have the same dimension of the GPT system. We do not adopt this approach. For example, it deems the rebit stabilizer theory to be nonclassical. However, the rebit stabilizer theory, in the prepare-and-measure scenario, admits of a noncontextual ontological model given by the Spekkens toy theory [70], as has been shown in several works [82–84].

While it would be arguably more appropriate to develop our hierarchy of contextuality for accessible fragments instead of plain GPT systems, we chose against this option for two reasons. Firstly, this would make the approach superficially more complicated, even though all the key ideas can be presented in the simpler framework of GPT systems. Secondly, one can easily accommodate our approach to the case of accessible fragments. The only difference is that the preparation (and outcome) equivalences have to be adjusted. For example, two preparations (i.e. distributions over $\tilde{\Omega}_{\mathsf{A}}$) would be deemed equivalent if their means cannot be distinguished by any effect in E_{A}.

On the resource theory of contextuality due to Duarte and Amaral. Another related work is [44], where the authors also provide a resource theory of generalized contextuality in prepare-and-measure scenarios using a 'black box' framework. In [44] the basic object consists of a *prepare-and-measure scenario* $\mathcal{S}$ together with a *behaviour* B. A scenario is a set $\mathcal{S} = \{\mathcal{P}, \mathcal{M}, \mathcal{D}, \mathcal{E}_P, \mathcal{E}_M\}$ where $\mathcal{P}$ is a set of preparations, $\mathcal{M}$ a set of measurements, $\mathcal{D}$ a set of outcomes, $\mathcal{E}_P$ the operational equivalences for preparations and $\mathcal{E}_M$ the operational equivalences for measurements.

A behaviour B is a set of conditional probabilities:

$$B = \{p(k|j,i)\}_{i\in\mathcal{I}, j\in\mathcal{J}, k\in\mathcal{K}} \tag{59}$$

corresponding to the probability of outcome d_k given preparation P_i and measurement M_j. $\mathcal{I}, \mathcal{J}$ and $\mathcal{K}$ are sets of labels for the preparations, measurements and outcomes, respectively.

We note that the sets of operational equivalences $\mathcal{E}_P$ and $\mathcal{E}_M$ of a given object are defined independently of the probability distributions. As such an object (scenario + behaviour) may contain operational equivalences which are not implied by the behaviour. Alternatively, there may be operational equivalences present in the behaviour which are not included in the scenario. In the cases where all ensembles of procedures are valid procedures in the scenario, and where the operational equivalences implied by behaviour are those of the scenario, then the object defined is equivalent to a GPT system.

The free objects of [44] are the noncontextual behaviours, namely those which admit of a generalized noncontextual model.

The operations of [44] are given by pre- and post-processing on the preparations, measurements and outcomes and correspond to channel simulations. The free operations correspond to channel simulations such that all the operational equivalences of the simulated system are images of operational equivalences of the simulating system through the simulation map.

Let us consider an example where both resource theories can be applied to highlight the differences between the resource theory of [44] and the one presented in this work.

We consider a scenario $\mathcal{S} = \{\mathcal{P}, \mathcal{M}, \mathcal{D}, \mathcal{E}_P, \mathcal{E}_M\}$ and a scenario $\tilde{\mathcal{S}} = \{\tilde{\mathcal{P}}, \tilde{\mathcal{M}}, \tilde{\mathcal{D}}, \tilde{\mathcal{E}}_P, \tilde{\mathcal{E}}_M\}$ where,

$$\mathcal{P} = \{0,1\}, \quad \mathcal{M} = \{0\}, \quad \mathcal{D} = \{0,1\}, \quad \mathcal{E}_P = \emptyset, \quad \mathcal{E}_M = \emptyset, \tag{60}$$

$$\mathcal{P} = \{0,1,2\}, \quad \mathcal{M} = \{0\}, \quad \mathcal{D} = \{0,1,2\}, \quad \mathcal{E}_P = \emptyset, \quad \mathcal{E}_M = \emptyset. \tag{61}$$

The behaviours on $\mathcal{S}$ and $\tilde{\mathcal{S}}$ are B and $\tilde{B}$ respectively:

$$p(k|j=1,i) = \delta_{i,k}, \ i,k \in \{0,1\}, \tag{62}$$

$$p(\tilde{k}|\tilde{j}=1,\tilde{i}) = \delta_{\tilde{i},\tilde{k}}, \ \tilde{i},\tilde{k} \in \{0,1,2\}. \tag{63}$$

These are the behaviours obtained by preparing a bit (resp. trit) in one of its extremal states and then implementing the canonical two outcome (resp. three outcome) measurement which returns one of the outcomes with certainty.

In the framework of GPTs $(\mathcal{S}, B)$ corresponds to a bit and $(\tilde{\mathcal{S}}, \tilde{B})$ to a trit (assuming one also allows all ensembles of preparations and measurements).

In the resource theory of GPT contextuality presented in this work we have that the bit and the trit are equivalent as resources: $\Delta_2 \succeq \Delta_3$ and $\Delta_3 \succeq \Delta_2$.

However in the resource theory of [44] this is not the case. We now show explicitly that there is no free operation T taking $(\mathcal{S}, B)$ to $(\tilde{\mathcal{S}}, \tilde{B})$. Since there is only one measurement in both scenarios

we drop the indices j and $\tilde{j}$. Let us assume the existence of stochastic maps $q_O(\tilde{k}|k)$ and $q_P(i|\tilde{i})$ which map $B \to \tilde{B}$:

$$p(\tilde{k}|\tilde{i}) = \sum_{i,k=0}^{1} q_O(\tilde{k}|k)p(k|i)q_P(i|\tilde{i}) \tag{64}$$

Now using the definitions of B and $\tilde{B}$ of Equation (62) we obtain:

$$\delta_{\tilde{i},\tilde{k}} = \sum_{i=0}^{1} q_O(\tilde{k}|i)q_P(i|\tilde{i}). \tag{65}$$

Let us denote by Q_O the 3×2 matrix with entries $[Q_O]_{\tilde{k},i} = q_O(\tilde{k}|i)$ and Q_P the 2×3 matrix with entries $[Q_P]_{i,\tilde{i}} = q_P(i|\tilde{i})$. Then the previous equation is equivalent to the matrix equation:

$$\mathbb{I}_3 = Q_O \cdot Q_P \tag{66}$$

This yields a contradiction since the decomposition of a 3×3 matrix A into matrix product $A = B \cdot C$ with B a 3×2 matrix and C a 2×3 matrix implies C has rank at most 2. However $\mathbb{I}_3$ is a rank 3 matrix. This shows that there is no free operation $(\mathcal{S}, B) \mapsto (\tilde{\mathcal{S}}, \tilde{B})$ in the resource theory of [44]. Notice that the reason why in our resource theory the bit and the trit — and any other noncontextual theory — are equivalent is that our free operations include the access to the classical system. This is the origin of the difference with the resource theory of [44].

An additional contribution of the present work is the following. In [44] the authors define a notion of composite of two behaviours which is equivalent to the minimal composite of Definition 5 in our framework. They then show that this composite is consistent with their resource theory, in the sense that the free objects are closed under this composite. In the present work we show that any consistent composite of GPT systems, not just the minimal composite of Definition 5 is compatible with the resource theory of contextuality.

6 Conclusion

Generalized contextuality, a leading notion of nonclassicality, is of crucial importance both in the foundations of quantum theory and in quantum information processing. Despite this, there is no complete characterisation of generalized contextuality as a resource. In this work we address this shortcoming.

Based on recent developments formulating noncontextuality of GPT systems via simplex embeddability, we have defined a resource theory of contextuality of GPT systems in prepare-and-measure scenarios. The free resources are the noncontextual systems and the free operations are univalent simulations with free access to classical systems. Using these notions, we motivate a hierarchy of contextuality for GPT systems. A new contextuality monotone arises naturally from our considerations — the classical excess, which expresses the minimum error of a univalent simulation by the countably-infinite classical system. We have also shown how a standard witness of contextuality, in the form of the POM success probability, can be used to define a contextuality monotone.

A key feature of the resource theory presented in this work is that the free operations include access to classical systems. Since every noncontextual system can be univalently simulated by a classical system this entails that all noncontextual systems are equivalent in the hierarchy. Without access to free classical systems the preorder which emerges is the embeddability preorder, for which generic noncontextual systems are inequivalent. Moreover we show, by assuming that the only way of composing GPT systems with classical ones is the standard one, that any consistent composition of GPT systems is consistent with the hierarchy: if $\mathsf{A}_1 \preceq \mathsf{B}_1$ and $\mathsf{A}_2 \preceq \mathsf{B}_2$ then $\mathsf{A}_1 \tilde{\otimes} \mathsf{A}_2 \preceq \mathsf{B}_1 \tilde{\otimes} \mathsf{B}_2$.

We have also discussed how GPT simulations could describe a physical process of information erasure that would explain the fine-tuning associated with contextuality[10] in a similar way to how the quantum equilibration process proposed by Valentini explains the fine-tuning associated with

nonlocality in Bohmian mechanics. We have argued that the information erasure in this case would be different than a simple coarse-graining. An interesting avenue for future research would be to further characterize this kind of information erasure and propose a test to detect heat dissipation in experiments manifesting contextuality. Even though such a proposal is undoubtedly radical, we believe that it could explain the fine-tuning associated with contextuality without abandoning the ontological models framework.

Another direction for further investigation is to develop a solid interpretation and establish the use-cases of the resource theory of GPT-contextuality that we provide. It is known that a resource theory of a given notion need not be unique. We have examples of this fact, like the resource theories of entanglement based on LOCC [86] and LOSR [87, 88] free operations, respectively. This does not mean that only one of them is "the correct resource theory of entanglement". Rather, these resource theories may be applicable in different contexts or for different purposes. In the case of entanglement, one can say that LOCC operations are relevant for communication tasks in the context of quantum internet and LOSR operations are relevant for the study of entanglement in Bell scenarios.

In this respect, we have discussed the relation of our work with other studies of generalized contextuality for GPTs in section 5.2. One that we did not mention is [76]. It would be interesting to find a relationship between our classical excess and the simulation cost of contextuality defined therein. Finally, it would be also interesting to extend the methods developed here to the realm of resource theories of more general fine-tunings [4], such as violations of time symmetry [89] and bounded ontological distinctness [90].

Acknowledgments

The authors thank the participants of the PIMan workshop – Orange (CA), March 2019 – where the idea of this project originated. In particular, Luke Burns and Justin Dressel, who were part of the initial discussions on the project. The authors further thank Rafael Wagner for insightful explanations regarding the approach of Duarte and Amaral [44]. LC thanks Farid Shahandeh and TGa thanks Markus Müller for helpful discussions. This project started when LC was supported by the Fetzer Franklin Fund of the John E. Fetzer Memorial Trust and by the Army Research Office (ARO) (Grant No. W911NF-18-1-0178). LC also acknowledges funding from the Einstein Research Unit "Perspectives of a Quantum Digital Transformation" and from the Horizon Europe project FoQaCiA, GA no.101070558. TGa acknowledges support from the Austrian Science Fund (FWF) via project P 33730-N. This research was partly funded by the Austrian Science Fund (FWF) 10.55776/PAT2839723. TGo acknowledges support from the Austrian Science Fund. This research was funded in whole or in part by the Austrian Science Fund (FWF) via the START Prize Y1261-N. This research was supported in part by Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported by the Government of Canada through the Department of Innovation, Science, and Economic Development, and by the Province of Ontario through the Ministry of Colleges and Universities.

For open access purposes, the authors have applied a CC BY public copyright license to any accepted manuscript version arising from this submission.

References

[1] David Schmid. Generalized noncontextuality. *Solstice of Foundations, ETH Zurich*, 2022. https://www.youtube.com/watch?v=M3qn3EHWdOg.

[10] A recent work that connects preparation contextuality and information erasure is [85], where the authors show that any ontological model reproducing the statistics of a sequential protocol involving incompatible projective measurements involves more information erasure than what operational quantum theory predicts. This fact is strictly related to the presence of preparation contextuality, as the same final quantum state in the protocol is represented by two different ontic distributions. This result can be seen as an example of fine-tuning of information erasure (i.e. more erasure at the ontological level than at the operational level), which, despite being interesting, differs from our idea of viewing contextuality as arising from a process of information erasure that explains the operational equivalences of distinct ontological representations.

[2] R. W. Spekkens. Contextuality for preparations, transformations, and unsharp measurements. *Phys. Rev. A*, 71:052108, May 2005. doi:10.1103/PhysRevA.71.052108.
[3] Robert Spekkens. The ontological identity of empirical indiscernibles: Leibniz's methodological principle and its significance in the work of einstein. *arXiv:1909.04628*, 2019. doi:https://doi.org/10.48550/arXiv.1909.04628.
[4] Lorenzo Catani and Matthew Leifer. A mathematical framework for operational fine tunings. *Quantum*, 7:948, March 2023. doi:10.22331/q-2023-03-16-948.
[5] Christopher Ferrie and Joseph Emerson. Frame representations of quantum mechanics and the necessity of negativity in quasi-probability representations. *Journal of Physics A: Mathematical and Theoretical*, 41(35):352001, jul 2008. doi:10.1088/1751-8113/41/35/352001.
[6] Robert W. Spekkens. Negativity and contextuality are equivalent notions of nonclassicality. *Phys. Rev. Lett.*, 101:020401, Jul 2008. doi:10.1103/PhysRevLett.101.020401.
[7] J. S. Bell. On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1(3):195–200, 1964. doi:10.1103/PhysicsPhysiqueFizika.1.195.
[8] Michael D. Mazurek, Matthew F. Pusey, Ravi Kunjwal, Kevin J. Resch, and Robert W. Spekkens. An experimental test of noncontextuality without unphysical idealizations. *Nature Communications*, 7(1):ncomms11780, 2016. doi:10.1038/ncomms11780.
[9] Michael D. Mazurek, Matthew F. Pusey, Kevin J. Resch, and Robert W. Spekkens. Experimentally bounding deviations from quantum theory in the landscape of generalized probabilistic theories. *PRX Quantum*, 2:020302, Apr 2021. doi:10.1103/PRXQuantum.2.020302.
[10] Robert W. Spekkens, D. H. Buzacott, A. J. Keehn, Ben Toner, and G. J. Pryde. Preparation Contextuality Powers Parity-Oblivious Multiplexing. *Phys. Rev. Lett.*, 102(1):010401, 2009. doi:10.1103/PhysRevLett.102.010401.
[11] Alley Hameedi, Armin Tavakoli, Breno Marques, and Mohamed Bourennane. Communication games reveal preparation contextuality. *Phys. Rev. Lett.*, 119:220402, Nov 2017. doi:10.1103/PhysRevLett.119.220402.
[12] David Schmid and Robert W. Spekkens. Contextual advantage for state discrimination. *Phys. Rev. X*, 8:011015, Feb 2018. doi:10.1103/PhysRevX.8.011015.
[13] Debashis Saha and Anubhav Chaturvedi. Preparation contextuality as an essential feature underlying quantum communication advantage. *Phys. Rev. A*, 100:022108, Aug 2019. doi:10.1103/PhysRevA.100.022108.
[14] Matteo Lostaglio and Gabriel Senno. Contextual advantage for state-dependent cloning. *Quantum*, 4:258, April 2020. doi:10.22331/q-2020-04-27-258.
[15] Matteo Lostaglio. Certifying quantum signatures in thermodynamics and metrology via contextuality of quantum linear response. *Phys. Rev. Lett.*, 125:230603, Dec 2020. doi:10.1103/PhysRevLett.125.230603.
[16] Shiv Akshar Yadavalli and Ravi Kunjwal. Contextuality in entanglement-assisted one-shot classical communication. *Quantum*, 6:839, October 2022. doi:10.22331/q-2022-10-13-839.
[17] Kieran Flatt, Hanwool Lee, Carles Roch I Carceller, Jonatan Bohr Brask, and Joonwoo Bae. Contextual advantages and certification for maximum-confidence discrimination. *PRX Quantum*, 3:030337, Sep 2022. doi:10.1103/PRXQuantum.3.030337.
[18] Carles Roch i Carceller, Kieran Flatt, Hanwool Lee, Joonwoo Bae, and Jonatan Bohr Brask. Quantum vs noncontextual semi-device-independent randomness certification. *Phys. Rev. Lett.*, 129:050501, Jul 2022. doi:10.1103/PhysRevLett.129.050501.
[19] Lorenzo Catani, Matthew Leifer, Giovanni Scala, David Schmid, and Robert W. Spekkens. What is nonclassical about uncertainty relations? *Phys. Rev. Lett.*, 129:240401, Dec 2022. doi:10.1103/PhysRevLett.129.240401.
[20] Rafael Wagner, Anita Camillini, and Ernesto F. Galvao. Coherence and contextuality in a mach-zehnder interferometer. *Quantum*, 8:1240, February 2024. doi:10.22331/q-2024-02-05-1240.
[21] Lorenzo Catani, Matthew Leifer, Giovanni Scala, David Schmid, and Robert W. Spekkens. Aspects of the phenomenology of interference that are genuinely nonclassical. *Phys. Rev. A*, 108:022207, Aug 2023. doi:10.1103/PhysRevA.108.022207.
[22] Bob Coecke, Tobias Fritz, and Robert W. Spekkens. A mathematical theory of resources. *Information and Computation*, 250:59–86, 2016. doi:https://doi.org/10.1016/j.ic.2016.02.008. Quantum Physics and Logic.

[23] Eric Chitambar and Gilad Gour. Quantum resource theories. *Rev. Mod. Phys.*, 91:025001, Apr 2019. doi:10.1103/RevModPhys.91.025001.
[24] Gilad Gour. Resources of the quantum world. *arXiv preprint arXiv:2402.05474*, 2024. doi:https://doi.org/10.48550/arXiv.2402.05474.
[25] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki. Quantum entanglement. *Rev. Mod. Phys.*, 81:865–942, Jun 2009. doi:10.1103/RevModPhys.81.865.
[26] Tomáš Gonda. Resource theories as quantale modules. *arXiv preprint arXiv:2112.02349*, 2021. doi:https://doi.org/10.48550/arXiv.2112.02349.
[27] Lucien Hardy. Quantum theory from five reasonable axioms. *arXiv:quant-ph/0101012*, 2001. doi:https://doi.org/10.48550/arXiv.quant-ph/0101012.
[28] Jonathan Barrett. Information processing in generalized probabilistic theories. *Phys. Rev. A*, 75:032304, Mar 2007. doi:10.1103/PhysRevA.75.032304.
[29] Peter Janotta and Haye Hinrichsen. Generalized probability theories: what determines the structure of quantum theory? *Journal of Physics A: Mathematical and Theoretical*, 47(32): 323001, jul 2014. doi:10.1088/1751-8113/47/32/323001.
[30] Martin Plávala. General probabilistic theories: An introduction. *Physics Reports*, 1033:1–64, 2023. doi:https://doi.org/10.1016/j.physrep.2023.09.001. General probabilistic theories: An introduction.
[31] Daniel Gottesman. Stabilizer codes and quantum error correction. *PhD thesis, California Institute of Technology*, 1997. doi:https://doi.org/10.48550/arXiv.quant-ph/9705052.
[32] Markus P. Müller and Andrew J. P. Garner. Testing quantum theory by generalizing noncontextuality. *Phys. Rev. X*, 13:041001, Oct 2023. doi:10.1103/PhysRevX.13.041001.
[33] Simon Kochen and E. P. Specker. *The Problem of Hidden Variables in Quantum Mechanics*, pages 293–328. Springer Netherlands, Dordrecht, 1975. doi:10.1007/978-94-010-1795-4_17.
[34] Samson Abramsky, Rui Soares Barbosa, and Shane Mansfield. Contextual fraction as a measure of contextuality. *Phys. Rev. Lett.*, 119:050504, Aug 2017. doi:10.1103/PhysRevLett.119.050504.
[35] Rui Soares Barbosa, Martti Karvonen, and Shane Mansfield. *Closing Bell Boxing Black Box Simulations in the Resource Theory of Contextuality*, pages 475–529. Springer International Publishing, Cham, 2023. doi:10.1007/978-3-031-24117-8_13.
[36] Martti Karvonen. Neither contextuality nor nonlocality admits catalysts. *Phys. Rev. Lett.*, 127:160402, Oct 2021. doi:10.1103/PhysRevLett.127.160402.
[37] Samson Abramsky and Adam Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, 13(11):113036, nov 2011. doi:10.1088/1367-2630/13/11/113036.
[38] Matthias Kleinmann, Otfried Gühne, José R Portillo, Jan-Åke Larsson, and Adán Cabello. Memory cost of quantum contextuality. *New Journal of Physics*, 13(11):113011, nov 2011. doi:10.1088/1367-2630/13/11/113011.
[39] Karl Svozil. How much contextuality? *Natural Computing*, 11(2):261–265, 2012. doi:10.1007/s11047-012-9318-9.
[40] A. Grudka, K. Horodecki, M. Horodecki, P. Horodecki, R. Horodecki, P. Joshi, W. Kłobus, and A. Wójcik. Quantifying contextuality. *Phys. Rev. Lett.*, 112:120401, Mar 2014. doi:10.1103/PhysRevLett.112.120401.
[41] Lu Li, Kaifeng Bu, and Junde Wu. Contextual robustness: An operational measure of contextuality. *Phys. Rev. A*, 101:012120, Jan 2020. doi:10.1103/PhysRevA.101.012120.
[42] Karol Horodecki, Jingfang Zhou, Maciej Stankiewicz, Roberto Salazar, Paweł Horodecki, Robert Raussendorf, Ryszard Horodecki, Ravishankar Ramanathan, and Emily Tyhurst. The rank of contextuality. *New Journal of Physics*, 25(7):073003, jul 2023. doi:10.1088/1367-2630/acdf78.
[43] Barbara Amaral. Resource theory of contextuality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377(2157):20190010, 2019. doi:10.1098/rsta.2019.0010.
[44] Cristhiano Duarte and Barbara Amaral. Resource theory of contextuality for arbitrary prepare-and-measure experiments. *Journal of Mathematical Physics*, 59(6):062202, 06 2018. doi:10.1063/1.5018582.

[45] David Schmid, Robert W. Spekkens, and Elie Wolfe. All the noncontextuality inequalities for arbitrary prepare-and-measure experiments with respect to any fixed set of operational equivalences. *Phys. Rev. A*, 97:062103, Jun 2018. doi:10.1103/PhysRevA.97.062103.

[46] Rafael Wagner, Roberto D Baldijão, Alisson Tezzin, and Bárbara Amaral. Using a resource theoretic perspective to witness and engineer quantum generalized contextuality for prepare-and-measure scenarios. *Journal of Physics A: Mathematical and Theoretical*, 56(50):505303, nov 2023. doi:10.1088/1751-8121/ad0bcc.

[47] Martin Plávala. Incompatibility in restricted operational theories: connecting contextuality and steering. *Journal of Physics A: Mathematical and Theoretical*, 55(17):174001, 2022.

[48] David Schmid, John H. Selby, Elie Wolfe, Ravi Kunjwal, and Robert W. Spekkens. Characterization of noncontextuality in the framework of generalized probabilistic theories. *PRX Quantum*, 2:010331, Feb 2021. doi:10.1103/PRXQuantum.2.010331.

[49] Victor Gitton and Mischa P. Woods. Solvable Criterion for the Contextuality of any Prepare-and-Measure Scenario. *Quantum*, 6:732, June 2022. doi:10.22331/q-2022-06-07-732.

[50] Victor Gitton and Mischa P. Woods. On the system loophole of generalized noncontextuality. *arXiv:2209.04469*, 2022. doi:https://doi.org/10.48550/arXiv.2209.04469.

[51] Farid Shahandeh. Contextuality of general probabilistic theories. *PRX Quantum*, 2:010330, Feb 2021. doi:10.1103/PRXQuantum.2.010330.

[52] John H. Selby, David Schmid, Elie Wolfe, Ana Belén Sainz, Ravi Kunjwal, and Robert W. Spekkens. Accessible fragments of generalized probabilistic theories, cone equivalence, and applications to witnessing nonclassicality. *Phys. Rev. A*, 107:062203, Jun 2023. doi:10.1103/PhysRevA.107.062203.

[53] John H. Selby, Elie Wolfe, David Schmid, Ana Belén Sainz, and Vinicius P. Rossi. Linear program for testing nonclassicality and an open-source implementation. *Phys. Rev. Lett.*, 132: 050202, Jan 2024. doi:10.1103/PhysRevLett.132.050202.

[54] Nicholas Harrigan and Robert W. Spekkens. Einstein, Incompleteness, and the Epistemic View of Quantum States. *Foundations of Physics*, 40(2):125–157, 2010. doi:10.1007/s10701-009-9347-0.

[55] George W. Mackey. *Mathematical Foundations of Quantum Mechanics*. A. Benjamin, Inc., New York, 1963.

[56] Gunther Ludwig. Versuch einer axiomatischen grundlegung der quantenmechanik und allgemeinerer physikalischer theorien. *Zeitschrift für Physik*, 181(3):233–260, Jun 1964. doi:10.1007/BF01418533.

[57] E. Brian Davies and John T. Lewis. An operational approach to quantum probability. *Communications in Mathematical Physics*, 17(3):239–260, 1970.

[58] Günther Ludwig. *An Axiomatic Basis for Quantum Mechanics: Volume 1 Derivation of Hilbert Space Structure*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985. OCLC: 858930098.

[59] Ludovico Lami. Non-classical correlations in quantum mechanics and beyond. *arXiv:1803.02902*, 2018. doi:https://doi.org/10.48550/arXiv.1803.02902.

[60] Roberto Beneduci and Leon Loveridge. Incompatibility of effects in general probabilistic models. 55(25):254005, may 2022. doi:10.1088/1751-8121/ac6f9d.

[61] Aleksandr S. Holevo. *Probabilistic and statistical aspects of quantum theory*. Number 1 in Quaderni Monographs. Edizioni della normale, Pisa, 2., english ed edition, 2011. OCLC: 746305136.

[62] E G Beltrametti and S Bugajski. A classical extension of quantum mechanics. *Journal of Physics A: Mathematical and General*, 28(12):3329, jun 1995. doi:10.1088/0305-4470/28/12/007.

[63] Bob Coecke, Tobias Fritz, and Robert W. Spekkens. A mathematical theory of resources. *Information and Computation*, 250:59–86, 2016. doi:https://doi.org/10.1016/j.ic.2016.02.008. Quantum Physics and Logic.

[64] Nicholas Gauguin Houghton-Larsen. A mathematical framework for causally structured dilations and its relation to quantum self-testing. *arXiv:2103.02302*, 2021. doi:https://doi.org/10.48550/arXiv.2103.02302.

[65] Guillaume Aubrun, Ludovico Lami, Carlos Palazuelos, and Martin Plávala. Entangleability of cones. *Geometric and Functional Analysis*, 31(2):181–205, April 2021. doi:10.1007/s00039-021-00565-5.

[66] Guillaume Aubrun, Ludovico Lami, Carlos Palazuelos, and Martin Plávala. Entanglement and superposition are equivalent concepts in any physical theory. *Phys. Rev. Lett.*, 128:160402, Apr 2022. doi:10.1103/PhysRevLett.128.160402.

[67] Giacomo Mauro D'Ariano, Marco Erba, and Paolo Perinotti. Classical theories with entanglement. *Phys. Rev. A*, 101:042118, Apr 2020. doi:10.1103/PhysRevA.101.042118.

[68] Tobias Fritz. Resource convertibility and ordered commutative monoids. *Mathematical Structures in Computer Science*, 27(6):850–938, 2017. doi:10.1017/S0960129515000444.

[69] Tomáš Gonda and Robert W. Spekkens. Monotones in General Resource Theories. *Compositionality*, 5, August 2023. doi:10.32408/compositionality-5-7.

[70] Robert W. Spekkens. Evidence for the epistemic view of quantum states: A toy theory. *Phys. Rev. A*, 75:032110, Mar 2007. doi:10.1103/PhysRevA.75.032110.

[71] Manik Banik, Some Sankar Bhattacharya, Amit Mukherjee, Arup Roy, Andris Ambainis, and Ashutosh Rai. Limited preparation contextuality in quantum theory and its relation to the cirel'son bound. *Phys. Rev. A*, 92:030103, Sep 2015. doi:10.1103/PhysRevA.92.030103.

[72] André Chailloux, Iordanis Kerenidis, Srijita Kundu, and Jamie Sikora. Optimal bounds for parity-oblivious random access codes. *New Journal of Physics*, 18(4):045003, apr 2016. doi:10.1088/1367-2630/18/4/045003.

[73] Shouvik Ghorai and A. K. Pan. Optimal quantum preparation contextuality in an n-bit parity-oblivious multiplexing task. *Phys. Rev. A*, 98:032110, Sep 2018. doi:10.1103/PhysRevA.98.032110.

[74] Debashis Saha, Paweł Horodecki, and Marcin Pawłowski. State independent contextuality advances one-way communication. *New Journal of Physics*, 21(9):093057, sep 2019. doi:10.1088/1367-2630/ab4149.

[75] Andris Ambainis, Manik Banik, Anubhav Chaturvedi, Dmitry Kravchenko, and Ashutosh Rai. Parity oblivious d-level random access codes and class of noncontextuality inequalities. *Quantum Information Processing*, 18(4):111, 2019. doi:10.1007/s11128-019-2228-3.

[76] Armin Tavakoli, Emmanuel Zambrini Cruzeiro, Roope Uola, and Alastair A. Abbott. Bounding and simulating contextual correlations in quantum theory. *PRX Quantum*, 2:020334, Jun 2021. doi:10.1103/PRXQuantum.2.020334.

[77] Lorenzo Catani, Ricardo Faleiro, Pierre-Emmanuel Emeriau, Shane Mansfield, and Anna Pappa. Connecting xor and xor* games. *Phys. Rev. A*, 109:012427, Jan 2024. doi:10.1103/PhysRevA.109.012427.

[78] Massy Khoshbin, Lorenzo Catani, and Matthew Leifer. Alternative robust ways of witnessing nonclassicality in the simplest scenario. *Phys. Rev. A*, 109:032212, Mar 2024. doi:10.1103/PhysRevA.109.032212.

[79] Alley Hameedi, Armin Tavakoli, Breno Marques, and Mohamed Bourennane. Communication games reveal preparation contextuality. *Phys. Rev. Lett.*, 119:220402, Nov 2017. doi:10.1103/PhysRevLett.119.220402.

[80] Antony Valentini. Signal-locality, uncertainty, and the subquantum h-theorem. i. *Physics Letters A*, 156(1):5 – 11, 1991. doi:https://doi.org/10.1016/0375-9601(91)90116-P.

[81] David Schmid, John H. Selby, and Robert W. Spekkens. Addressing some common objections to generalized noncontextuality. *Phys. Rev. A*, 109:022228, Feb 2024. doi:10.1103/PhysRevA.109.022228.

[82] Robert W. Spekkens. *Quasi-Quantization: Classical Statistical Theories with an Epistemic Restriction*, pages 83–135. Springer Netherlands, Dordrecht, 2016. doi:10.1007/978-94-017-7303-4_4.

[83] Lorenzo Catani and Dan E Browne. Spekkens' toy model in all dimensions and its relationship with stabiliser quantum mechanics. *New Journal of Physics*, 19(7):073035, jul 2017. doi:10.1088/1367-2630/aa781c.

[84] Lorenzo Catani and Dan E. Browne. State-injection schemes of quantum computation in spekkens' toy theory. *Phys. Rev. A*, 98:052108, Nov 2018. doi:10.1103/PhysRevA.98.052108.

[85] Alberto Montina and Stefan Wolf. Realism and causality imply information erasure by measurements. *arXiv:2307.03134*, 2023. doi:https://doi.org/10.48550/arXiv.2307.03134.

[86] Charles H. Bennett, Herbert J. Bernstein, Sandu Popescu, and Benjamin Schumacher. Concentrating partial entanglement by local operations. *Phys. Rev. A*, 53:2046–2052, Apr 1996. doi:10.1103/PhysRevA.53.2046.

[87] Francesco Buscemi. All entangled quantum states are nonlocal. *Phys. Rev. Lett.*, 108:200401, May 2012. doi:10.1103/PhysRevLett.108.200401.

[88] David Schmid, Thomas C. Fraser, Ravi Kunjwal, Ana Belen Sainz, Elie Wolfe, and Robert W. Spekkens. Understanding the interplay of entanglement and nonlocality: motivating and developing a new branch of entanglement theory. *Quantum*, 7:1194, December 2023. doi:10.22331/q-2023-12-04-1194.

[89] Matthew S. Leifer and Matthew F. Pusey. Is a time symmetric interpretation of quantum theory possible without retrocausality? *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2202):20160607, 2017. doi:10.1098/rspa.2016.0607.

[90] Anubhav Chaturvedi and Debashis Saha. Quantum prescriptions are more ontologically distinct than they are operationally distinguishable. *Quantum*, 4:345, October 2020. doi:10.22331/q-2020-10-21-345.

A On the physicality of simulations

In the framework of GPTs a physical map $\mathsf{B} \to \mathsf{A}$ is given by a linear map $M : V_\mathsf{B} \to V_\mathsf{A}$ satisfying $M(\Omega_\mathsf{B}) \subseteq \Omega_\mathsf{A}$. M can be viewed as part of a preparation procedure, preparing a state of A by first preparing a state ω_B of B and then applying M to obtain ω_A given by $M(\omega_\mathsf{B})$. Composing this with a measurement $\{e^i_\mathsf{A}\}_i$ on A gives outcome probabilities $e^i_\mathsf{A} \cdot M(\omega_\mathsf{B})$. However, one could also view this experiment as preparing the state ω_B on system B and then measuring $\{e^i_\mathsf{A} \circ M\}_i = \{M^* e^i_\mathsf{A}\}_i$ which is a measurement on B, hence we require that a physical map also obeys $M^*(e_\mathsf{A}) \in E_\mathsf{B}$ for all $e_\mathsf{A} \in E_\mathsf{A}$.

Definition 23 (Physical GPT transformation)**.** A physical GPT transformation M between two GPT systems, from $\mathsf{B} = (\Omega_\mathsf{B}, E_\mathsf{B}, V_\mathsf{B})$ to $\mathsf{A} = (\Omega_\mathsf{A}, E_\mathsf{A}, V_\mathsf{A})$, is given by a linear map $M : V_\mathsf{B} \to V_\mathsf{A}$ satisfying

$$M(\Omega_\mathsf{B}) \subseteq \Omega_\mathsf{A} \qquad \text{and} \qquad M^*(E_\mathsf{A}) \subseteq E_\mathsf{B}. \tag{67}$$

One may wonder whether an exact simulation (Γ, Θ) of a system A by a system B can correspond to a physical transformation $M : \mathsf{B} \to \mathsf{A}$,[11] in the sense that the state simulation Γ assigns, to each state $\omega \in \Omega_\mathsf{A}$, a subset of those states of B that could give rise to ω when M is applied.

Definition 24 (Physical realisation of a simulation)**.** Consider a simulation (Γ, Θ) of a GPT system A by a GPT system B and a physical GPT transformation $M : \mathsf{B} \to \mathsf{A}$. We say that M is a **physical realisation** of (Γ, Θ) iff we have $\Gamma(\omega) \subseteq M^{-1}(\omega)$ for all $\omega \in \Omega_\mathsf{A}$ and $\Theta(e) \supseteq \{M^*(e)\}$ for all $e \in E_\mathsf{A}$.[12]

For example, a classical GPT system Δ_n is embeddable within an n-level quantum GPT system Q_n. To describe the univalent simulation, consider a basis of the latter's Hilbert space, whose elements are denoted by $|i\rangle$ for i ranging from 1 to n. We can let the state simulation map each extremal point δ_i of the classical state space to the quantum state $|i\rangle\langle i|$. The effect simulation is then uniquely determined, it must send each δ_i^* to the effect $\rho \mapsto \langle i| \rho |i\rangle$. There is a physical realisation of this simulation given by the completely dephasing map $\mathsf{Q}_n \to \Delta_n$ which acts as

$$\rho \mapsto \sum_{i=1}^{n} \langle i| \rho |i\rangle \, \delta_i. \tag{68}$$

Requiring a simulation (Γ, Θ) of a GPT system A by a system B to admit of a physical realisation is a very strong condition. It has the following consequences.

[11] There is another possibility: namely, the simulation may correspond to a physical transformation $W : \mathsf{A} \to \mathsf{B}$, in the sense that the state simulation Γ assigns, to each state $\omega \in \Omega_\mathsf{A}$, the set of those states of B that could give rise to ω when M is applied. We do not treat this case here.

[12] Here, $M^{-1}(\omega)$ refers to the preimage of ω under M and $\{M^*(e)\}$ refers to the singleton set whose element is $M^*(e)$.

Proposition 25. *Consider a simulation* (Γ, Θ) *of a system* A *by a system* B *with a physical realisation* $M : \mathsf{B} \to \mathsf{A}$.

1. *M is surjective on states, i.e. we have* $M(\Omega_\mathsf{B}) = \Omega_\mathsf{A}$.

2. *If* $\dim(V_A) = \dim(V_B)$ *holds, then* (Γ, Θ) *is a univalent simulation and M is an isomorphism.*

Proof. By Definition 24, we have $\Gamma(\omega) \subseteq M^{-1}(\omega)$ for all $\omega \in \Omega_\mathsf{A}$. In particular, the preimage of every such ω under M must be non-empty, which means $M(\Omega_\mathsf{B}) \supseteq \Omega_\mathsf{A}$. Combining this with the first condition in (67) gives $M(\Omega_\mathsf{B}) = \Omega_\mathsf{A}$.

Let us now assume that the underlying vector spaces of A and B have the same dimension given by $d \in \mathbb{N}$. We prove that Γ is single-valued by contradiction. To this end, assume that there are two distinct $\gamma_1, \gamma_2 \in \Gamma(\omega)$ for some $\omega \in \Omega_\mathsf{A}$. By empirical adequacy of the simulation, we must have

$$f \cdot \gamma_1 = f \cdot \gamma_2 \tag{69}$$

for all $f \in \Theta(E_\mathsf{A})$. In particular, this means that the span of $\Theta(E_\mathsf{A})$ has dimension strictly less than d and no effect in $\Theta(E_\mathsf{A})$ can distinguish two states of Ω_B that differ by a scalar multiple of $\gamma_1 - \gamma_2$.

By the fact that effects of A can be distinguished by its states, and the fact that Ω_A is a convex set (not containing the zero vector), there are d states $\rho_i \in \Omega_\mathsf{A}$ whose span is the whole of V_A. To distinguish these d states, one needs at least d linearly independent effects, let us denote a choice of these by $e_i \in E_\mathsf{A}$. By empirical adequacy, any choice of d states $\sigma_i \in \Gamma(\rho_i)$ of B must also span V_B and be distinguishable by effects in

$$\Theta\left(\{e_i\}_{i=1}^d\right) \subseteq \Theta(E_\mathsf{A}). \tag{70}$$

This is now a contradiction because $\Theta(E_\mathsf{A})$, as we established earlier, has dimension strictly less than d. In conclusion, Γ must be single-valued. An analogous argument can be used to show that Θ must also be single-valued.

By the univalence of the simulation, we can thus think of Γ and Θ as convex-linear maps $\Omega_\mathsf{A} \to \Omega_\mathsf{B}$ and $E_\mathsf{A} \to E_\mathsf{B}$ respectively. Since $\operatorname{span}(\Omega_\mathsf{A}) = V_\mathsf{A}$ and $0 \notin \Omega_\mathsf{A}$ (by the second condition in (5)) it follows that we can actually identify the state simulation with a linear map $V_\mathsf{A} \to V_\mathsf{B}$ denoted by $\bar{\Gamma}$.

Since Γ is injective by Implication (20) (which implies that $\bar{\Gamma}$ is as well) and we have $V_\mathsf{A} \cong V_\mathsf{B}$ by assumption, $\bar{\Gamma}$ is an isomorphism.

By Definition 24, we then have $M\Gamma(\omega) = \omega$ for all $\omega \in \Omega_\mathsf{A}$. By linearity, we can extend this equation to the span of Ω_A to get $M\bar{\Gamma} = \mathrm{id}_{V_\mathsf{A}}$, which means that M is the inverse of $\bar{\Gamma}$ since the latter is an isomorphism. Thus, M itself must be an isomorphism. □

In the case of unequal dimensions of V_A and V_B, a simulation can indeed be realised also by a physical GPT transformation that is not an isomorphism. One example is the simulation of the n-level classical system Δ_n by the quantum one Q_n mentioned above. The completely dephasing map from (68) gives a realisation thereof.

One of the consequences of Proposition 25 is that whenever there exists a simulation of a GPT system A by B and $\dim(V_\mathsf{A}) = \dim(V_\mathsf{B})$ holds, the state spaces of A and B must be isomorphic as convex sets. However, the same is not true for the effect spaces, as the following example shows.

Example 26 (Physical realisation of univalent simulation for inequivalent effect spaces). Consider two systems A and $\bar{\mathsf{A}}$ with identical underlying vector spaces, where we also have $\Omega_\mathsf{A} = \Omega_{\bar{A}}$ and $E_{\bar{A}} \subsetneq E_\mathsf{A}$. Then the inclusion maps on states and effects provide a univalent simulation of $\bar{\mathsf{A}}$ by A. Moreover, it is realisable by the physical GPT transformation given by the identity $\mathbb{I} : V_\mathsf{A} \to V_{\bar{\mathsf{A}}}$, since we have $\mathbb{I}(\Omega_\mathsf{A}) = \Omega_{\bar{\mathsf{A}}}$ and $\mathbb{I}^*(E_{\bar{\mathsf{A}}}) \subseteq E_\mathsf{A}$. Even though the identity is invertible as a linear map, its inverse $\mathbb{I} : V_{\bar{\mathsf{A}}} \to V_\mathsf{A}$ is not a physical GPT transformation since $\mathbb{I}^*(E_\mathsf{A}) \subseteq E_{\bar{\mathsf{A}}}$ is false by assumption.

We can apply Proposition 25 to the (convexified) Spekkens toy bit T_2 (with six extremal states), which has a non-contextual ontological model, i.e. it can be simulated by a classical GPT system, in particular by Δ_4 with four extremal states. Since these two GPT systems have isomorphic

underlying vector spaces and non-isomorphic state spaces, it follows that there can be no physical realisation $\Delta_4 \to \mathsf{T}_2$ of the univalent simulation of T_2 by Δ_4.

Interestingly, for any GPT system A with finitely many extremal states there is an ontological model thereof, which admits of a physical realisation.

Lemma 27. *The HBB simulation (see Example 9) of a GPT system A with $n \in \mathbb{N}$ extremal states by the classical system Δ_n has a physical realisation.*

Proof. Given a GPT system A with a finite set $\{\omega_i\}_{i=1}^n$ of extremal states, the HBB model is given by the maps (Γ, Θ) as defined in Example 9. We first show that the opposite $\Gamma^{-1} : \Gamma(\Omega_\mathsf{A}) \to \Omega_\mathsf{A}$ of Γ (which is a function as shown in (20)) uniquely extends to a linear map $M : \mathbb{R}^n \to V_\mathsf{A}$.

The map $\Gamma : \Omega_\mathsf{A} \to \Delta_n$ is a multivalued function defined as:

$$\Gamma(\omega) = \left\{ \sum_{i=1}^n p_i \delta_i \;\middle|\; p_i \in [0,1],\, \sum_{i=1}^n p_i = 1,\, \omega = \sum_{i=1}^n p_i \omega_i \right\}. \tag{71}$$

For an arbitrary element $\delta \in \Delta_n$, which can be uniquely decomposed as

$$\delta = \sum_i p_i \delta_i, \tag{72}$$

the function Γ^{-1} is given by

$$\Gamma^{-1}(\delta) = \sum_i p_i \omega_i \tag{73}$$

and is in particular a convex-linear map $\Delta_n \to \Omega_\mathsf{A}$.

Since Δ_n spans $\mathbb{R}^n$ and does not contain the origin, the map Γ^{-1} uniquely extends to a linear map $M : \mathbb{R}^n \to V_\mathsf{A}$ satisfying $M(\Delta_n) = \Omega_\mathsf{A}$ and $\Gamma(\omega) \subseteq M^{-1}(\omega)$ for all $\omega \in \Omega_\mathsf{A}$.

This map also satisfies $M^*(E_\mathsf{A}) \subseteq \Delta_n^*$, since for any $e \in E_\mathsf{A}$ we have

$$\forall \delta \in \Delta_n \qquad M^*(e) \cdot \delta = e \cdot M(\delta) \in [0,1], \tag{74}$$

so that $M^*(e)$ is an element of Δ_n^*. Therefore, M is a physical GPT transformation.

The univalent effect simulation map is given (as a function) by $\Theta(e) = \sum_i (e \cdot \omega_i) \delta_i^*$. We now have, for any $j \in \{1, 2, \ldots, n\}$,

$$M^*(e) \cdot \delta_j = e \cdot M(\delta_j) = e \cdot \omega_j = \sum_i (e \cdot \omega_i)(\delta_i^* \cdot \delta_j) = \Theta(e) \cdot \delta_j, \tag{75}$$

so that $M^*(e_\mathsf{B}) = \Theta(e_\mathsf{B})$ holds by (8). In conclusion, this shows that M is indeed a physical realisation of the HBB simulation. □