Perturbative Causality

Alexander S. Blum*, James D. Fraser†

*Max Planck Institute for the History of Science and Albert Einstein Institute, Potsdam
†IHPST, CNRS and Paris 1-Panthéon-Sorbonne University

Abstract

This paper examines the development of causal perturbation theory, a reformulation of perturbative quantum field theory (QFT) starting from a causality condition rather than a time-evolution equation. We situate this program alongside other causality-based reformulations of relativistic quantum theory which flourished in the post-war period, contrasting it in particular with axiomatic QFT. Whereas the axiomatic QFT tradition tried to move beyond the perturbative expansion, causal perturbation theory can be thought of as a foundational investigation of this approximation method itself. Unearthing this largely forgotten research program helps clarify questions of contemporary philosophical interest, for instance about the interpretative significance of the ultraviolet divergences which appear in the series expansion, but also help us understand why causality conditions became so ubiquitous in post-war high-energy theory.

1 Introduction

The 1950s and 1960s were crucial yet tumultuous decades in the development of relativistic quantum theory. While the invention of renormalized perturbative theory in the late 40s had been a victory of sorts, as time went on it was increasingly felt that something quite different was needed to resolve the theoretical and empirical problems posed by nuclear interactions. Accordingly, we see a proliferation of new ideas and research programs in this period, some of which may be familiar today—axiomatic QFT and the renormalization group—others perhaps less so—the theory of dispersion relations and the analytic S-matrix. Despite this diversity, one also finds an intriguing consensus amongst many of the programs that flourished in this period about the fundamental importance of some notion of causality to the formulation of relativistic quantum theory, though exactly what this meant was, as we shall see, variously conceived.

How did causality become so central to the debate about the formulation of relativistic quantum theory, and what does this tell us about the structure of contemporary QFTs? This paper focuses on a lesser known strand of 1950s high energy theory, namely the development of the so-called causal perturbation theory program, which we claim sheds light on these questions. The central issue driving the development of causal perturbation theory was the question of how the perturbative approximation scheme ought to be derived. Freeman Dyson's formulation of perturbative QFT had been based on the integration of the Schwinger-Tomonaga evolution equation. Ernst Stueckelberg, Nicolay Bogoliubov and their collaborators developed an alternative derivation of the series expansion starting not from a differential evolution equation but from a causality condition. This would eventually lead to a novel perspective on the problem of ultraviolet divergences and perturbative renormalization which remains relevant (and neglected) in contemporary interpretative debates.

Notions of causality remain a central point of concern in contemporary debates about the foundations of QFT and have been discussed extensively in the philosophy of physics literature (see for instance, Earman and Valente (2014); Calderón (2024)). Much of this work has focused on the various causality axioms adopted in algegraic axiomatizations of QFT leaving aside the question of why alternative definitions of causality became so significant in debates about the formulation of relativistic quantum theories in the first place. The analysis of this paper highlights how concerns about causality were integral to the reception and refinement of renormalized perturbation theory, as well as being operative in attempts to develop other non-perturbative approximation methods capable of describing strongly interacting nuclear phenomena. The causality condition which Bogoliubov eventually landed on as the foundation of the causal perturbation theory approach is also quite different in character from the microcausality condition typically appealed to in non-perturbative axiomatizations of QFT, raising questions about the relationships between these causality notions which are yet to be explored in the philosophical literature.

The paper is structured as follows. Sections 2 and 3 cover necessary background, introducing Heisenberg's S-matrix program and Freeman Dyson's derivation of the QFT perturbative expansion respectively and the critiques which both of these approaches to relativistic quantum theory faced. Sections 4-6 trace the development of the causal perturbation theory approach, with a particular emphasis on the struggle to articulate an appropriate causality condition. Sections 7-9 reflect on the implications of the causal derivation of the perturbative expansion: section 7 discusses the novel treatment of the ultraviolet divergences problem which develops within the causal perturbative theory approach; section 8 discusses the relationship between causal perturbation theory and more ambitious non-perturbative programs like axiomatic QFT and the theory of dispersion relations; section 9 concludes with some general morals for philosophical engagement with

causality concepts in relativistic quantum theory.

2 The Legacy of Heisenberg's S-matrix Program

From the early days of quantum mechanics, it was clear that combining the new framework with special relativity was a highly non-trivial task. Quantizing a classical field theory and using a perturbative approximation scheme to treat the concomitant non-linear interacting field equations emerged as perhaps the most promising route to a quantitatively predictive formalism. Following this recipe, the Heisenberg and Pauli (1929) formulation of QED was, in many respects, already very similar to the modern perturbative treatment of that theory. At the time, however, Heisenberg-Pauli QED was widely viewed as an impoverished (perhaps even mathematically inconsistent) stepping stone on the road to a more complete theory. There were at least two reasons for this negative assessment.

The first was a difficulty with representing the dynamics of a QFT in a relativistically covariant way. Heisenberg and Pauli based their formalism on the so-called equal-time commutation relations of the field operators, which distinguished the time and space arguments of the fields and was therefore not manifestly relativistic. As a result, a notoriously laboured argument was needed to demonstrate the covariance of the full theory. Another way to see the difficulty with covariance is to look at the Schrödinger equation, which was still taken to govern the Schrödinger picture state evolution in a field theoretic context:

$$id/dt|\psi(t)\rangle = H|\psi(t)\rangle \tag{1}$$

One issue with this equation is that the Schrödinger picture Hamiltonian is not a Lorentz scalar. Perhaps more fundamentally, in singling out the time coordinate, the Schrödinger equation required selecting a foliation of space-time in order to implement the dynamics. These difficulties with covariance all seemed to stem from the role of the Hamiltonian formalism in the canonical approach to quantization, leading many theorists in this period to try to generalize or replace it—including Heisenberg, as we shall see.[1]

The second, and apparently more cataclysmic, problem with Heisenberg-Pauli QED was the appearance of ultraviolet divergences in its perturbative expansions. As the prospects of exactly solving an interacting QFT seemed quite hopeless Heisenberg and Pauli adopted the now familiar strategy of expanding quantities in a series expansion in the interaction coupling. If the interaction was weak then the leading terms of this series ought to provide a good approximation of the relevant quantities in the fully interacting model. It quickly

[1]See Tomonaga (1966) for a discussion of covariance worries in the 1930s and 1940s, including the ideas of Dirac and Yukawa that apparently stimulated his work on the Schwinger-Tomonaga equation (discussed further in section 3).

became evident that, from second-order, the coefficients of this expansion contained integrals which diverged at coincident space-time points, or in momentum space as the momentum variable becomes infinitely large. There was clearly something deeply wrong with the Heisenberg-Pauli scheme, though exactly what was unclear. Were the divergences a manifestation of some deep inconsistency in the theoretical principles of QFT? Or could they be circumvented by either abandoning the perturbative approximation scheme or somehow mending it?

The divergence and covariance issues appeared to many theorists at the time to be deeply entwined with one another. On the other hand, the difficulty with maintaining covariance was seen as a major contributor to the intractability of the ultraviolet divergence problem, a view expressed clearly by Oppenheimer in his contribution to the 1948 Solvay conference:

> "[O]ne needs a covariant way of identifying these [divergent] terms; and for that, not merely the field equations themselves, but the whole method of approximation and solution must at all stages preserve covariance."

From this perspective, the difficulty with maintaining covariance made it hard to assess whether the ultraviolet divergence problem was truly fatal or could somehow be worked around. On the other hand, naive attempts to modify the basic principles of QFT to address the ultraviolet divergences problem seemed to make the clash with relativity considerably worse. Heisenberg came to see the ultraviolet divergence problem as stemming from the use of a differential time-evolution equation to express the theories dynamics, since this necessitated the multiplication of field operators at the same space-time point—the ultimate source of the divergences. He therefore came to the view that a future theory should incorporate a minimal length scale which would cut off the divergent integrals. Introducing a minimal length scale directly—for instance by discretizing the field equations—inevitably explicitly violated Lorentz symmetry, however.[2]

Heisenberg's response to this seemingly inevitable tension was to propose an entirely new dynamical framework for relativistic quantum mechanics, which eschewed the use of a differential time-evolution equation entirely (Heisenberg 1943a,b, 1944). To this end, he introduced the S-matrix, an operator which maps asymptotic states at $t = -\infty$, interpreted as 'incoming' states prior to a scattering process, to 'outgoing' asymptotic states at $t = \infty$, thus encoding the scattering cross section observables typically measured at scattering experiments. Heisenberg's hope was that it might be possible to get rid of the field operators, along with the usual method of canonical quantization with its reliance

[2]See Carazza and Kragh (1995) for a discussion of Heisenberg's early attempts to formulate a discrete quantum theory and Blum (2017) the later development of Heisenberg's views about the presence of a fundamental length.

on the Hamiltonian formalism, and construct models directly by imposing conditions on the S-matrix, the two he was able to come up with being unitarity and Lorentz invariance:

$$\begin{array}{ll} \textbf{Unitarity} & S^\dagger S = \mathbb{I}, \\ \textbf{Lorentz Invariance} & U_\Lambda S U_\Lambda^{-1} = S, \end{array}$$

where U_Λ are Lorentz transformations. It quickly became clear that these two conditions were insufficient to extract any quantitative information on their own, however; Heisenberg's bold new formalism was too austere to be practically useful.

A naive folk history of the period has it that Heisenberg's S-matrix program was essentially undercut by the work of Feynman, Schwinger, Tomonaga and Dyson. These authors showed that a relativistic evolution equation could in fact be formulated for QED and that the ultraviolet divergences could be systematically removed from the perturbative expansion via a procedure which came to be known as renormalization, thus resolving the two fundamental problems with Heisenberg-Pauli QED. Dyson played a key role in synthesising these insights, and showing that the new renormalized perturbative expansion was actually an expansion of Heisenberg's S-matrix, now viewed as a derived rather than fundamental object. The time-evolution equation retook its place as the core expression of the theory's dynamics, and Heisenberg's S-matrix program was largely forgotten, briefly resurfacing as an inspiration for Geoffrey Chew's analytic S-matrix program in the 1960s, before fading into obscurity once more.

This narrative is misleading in our view, however. In reality, Heisenberg's S-matrix program remained influential after the empirical success of renormalized perturbative QED and we can see a number of theoretical approaches which appeared in the 1950s as continuations of it—including the causal perturbation theory program. The reason for this resurgence of S-matrix ideas was that Dyson's formulation of renormalized perturbation theory eventually came to be seen—ironically, like Heisenberg-Pauli QED before it—as an incomplete stepping stone to a more satisfactory formulation of relativistic quantum theory. By the mid 1950s there was a sense that, while Heisenberg's principles of unitarity and Lorentz invariance were too weak, in attempting to rehabilitate the time-evolution equation Dyson's formalism had added back too much additional structure. This is where causality entered the picture. Adding some form of causality condition to unitarity and Lorentz invariance was a way to try to chart a course between Heisenberg and Dyson. As we shall discuss further in section 8, causal perturbation theory, axiomatic QFT and the theory of dispersion relations can all be read as pursuing variations of this general strategy.

Before we can get back to this synoptic view of these various causality based reformulation programs we intend to examine the historical origins of causal perturbation theory in

detail. Our starting point, in the following section, will be a more detailed examination of Dyson's derivation of the renormalized perturbative expansion, with a particular focus on how causality properties appear in that formalism.

3 Dysonian Perturbation Theory

As has often been remarked, the work of Feynman, Schwinger, Tomonaga and Dyson was theoretically conservative, in the sense that it did not modify any of the basic theoretical principles which had been used to formulate Pauli-Heisenberg QED. Indeed, it even maintained the perturbative approximation strategy which Pauli and Heisenberg had used to try to articulate the content of their theory. What "the men who made it" really made was an improved perturbative approximation scheme. This section presents the derivation of the new renormalized perturbation series found in Dyson (1949a,b), which we shall refer to as Dysonian perturbation theory for short. While Dyson's derivation remains influential today, a dilligent student of QFT will note that contemporary textbooks often take a different route to the series expansion of the S-matrix, starting from the path integral expression for the partition function and proceeding via the LSZ reduction formula. During our period of interest, however, Dyson's papers provided the canonical formulation of renormalized perturbative QED and represent the foil against which causal perturbation theory developed. Accordingly, this section will highlight the status of causality properties in Dyson's approach, preparing the way for our investigation of the alternative causality condition based derivation of the expansion developed by Stueckelberg and Bogoliubov in sections 4-6. How the path integral derivation of renormalized perturbation theory plays into to the historical and conceptual issues examined in this paper is left as an open question for future research.[3]

Dyson's derivation of a series expansion for the S-matrix starts from the Schwinger-Tomonaga equation—one answer to the worries about manifest covariance discussed in section 2. A key innovation here was the so-called interaction picture. In this representation of the time-evolution, the Hamiltonian of a field theory is split into a free and interacting part, $H = H_0 + H_I$; roughly speaking, H_0 is taken to describe the asymptotic (presumed free particle) in and out states, while H_I is taken to describe the dynamics of the

[3]Here are a few preliminary comments about the path integral derivation of the perturbative expansion. The basic form of this derivation can, in fact, be found in the appendix of Lehmann et al. (1955), so this derivation also goes back to the 1950s. Note, however, that in contrast to the causal perturbation theory derivation, the LSZ derivation was not conceived of as an improvement on Dysonian perturbation theory; rather the point was to demonstrate that perturbative QFT could be recovered within their non-perturbative formulation of QFT. The path integral formulation of perturbation theory seems only to have become more widely known with the rise of non-abelian gauge theories and the motivation for its adoption in this context was arguably calculational convenience rather than foundational.

scattering process. It is thus only H_I that appears in the interaction-picture Schrodinger equation

$$id/dt|\psi(t)\rangle = H_I|\psi(t)\rangle, \tag{2}$$

with the remaining (free) time evolution shifted into the operators. H_I is a Lorentz scalar so one issue with the non-covariance of the time-evolution equation is immediately addressed in the interaction picture. However, equation (2) still singles out of the time coordinate and thus requires the adoption of a particular foliation of space-time. Tomonaga (1946) argued that in the interaction picture it was possible to write down a fully relativistic analogue of the Schrödinger equation now known as the Schwinger-Tomonaga equation, which treats space and time on an equal footing:

$$i\frac{\delta\psi(\sigma)}{\delta\sigma(x)} = \mathcal{H}_I(x)\psi(\sigma). \tag{3}$$

Here $\sigma(x)$ is a space-like Cauchy surface containing the point x, and $\mathcal{H}_I(x)$ is the interaction Hamiltonian density (the interaction Hamiltonian H_I being equal to this density integrated over all space).

Dyson's derivation of the perturbative expansion proceeds via an integration of this equation. We would like to highlight two aspects of the dynamical framework adopted by Dyson: firstly, how a microcausality condition is incorporated into the covariant Schwinger-Tomonaga framework, and secondly, how the iterative integration of the time-evolution equation automatically leads to the time-ordering of operators appearing in the perturbative expansion coefficients.

The property of microcausality can be viewed as a necessary precondition for the cogency of the Schwinger-Tomonaga equation. One way to read (3) is as a system of local equations describing the time evolution of the quantum state at each point $\mathbf{x}$ in space. Adopting a particular foliation, we can split the point x into a spatial coordinate $\mathbf{x}$ and a time t; the variational derivative then simply becomes a time derivative and the Schwinger-Tomonaga equation describes the evolution of the quantum state at $\mathbf{x}$. For this to work, however, the time evolution at all points $\mathbf{x}$ needs to be independent, i.e., the Hamiltonian densities that generate the local time evolutions need to commute with one another. While we had to adopt a foliation to read the Schwinger-Tomonaga equation in this way, this should work for an arbitrary choice of foliation, so the requirement becomes:

$$[\mathcal{H}_I(x), \mathcal{H}_I(y)] = 0, \text{ if } (x-y)^2 < 0. \tag{4}$$

This is a special case of what is nowadays commonly called the microcausality condition: the requirement that operators associated with space-like separated regions commute. As Pauli, who first introduced this property in his original proof of the spin-statistics theorem, wrote:

> The justification for our postulate lies in the fact that measurements at two space points with space-like distance can never disturb each other, since no signals can be transmitted with velocities greater than that of light. (Pauli 1940, 721)

In other words, the "non-causal" behaviour that the microcausality condition is designed to rule out is superluminal influence. This is the first of two causality notions we will distinguish in this section.

Following the above line of reasoning, Tomonaga (1946) takes microcausality to be a necessary condition for the integrability of the Schwinger-Tomonaga equation. Furthermore, Dyson (1949b) appeals to the microcausality of the interaction Hamiltonian density explicitly in order to transition from the functional equation (3) to an ordinary differential evolution equation describing infinitesimal transitions between preselected space-like Cauchy surfaces. It is this foliated equation which Dyson actually integrates to obtain the series expansion of the S-matrix. Indeed, in later years it was viewed as superfluous to start from the covariant Schwinger-Tomonaga equation at all: one obtains exactly the same expression for the S-Matrix by adopting a particular foliation and starting from an interaction picture Schrodinger equation. Bjorken and Drell (1964) remark when presenting the Dysonian derivation of the perturbation series that the time evolution "may be covariantly defined on a general space-like surface instead of at constant t, but to no great advantage" and later textbook presentation followed their lead on this point. Maintaining covariance in particular contexts, such as the treatment of longitudinal and transverse components of the electromagnetic field, did prove to be crucial for the renormalization of QED perturbation theory, but Oppenheimer's statement about maintaining covariance "at all stages" in order to cure the ultraviolet divergences problem quoted above, while an accurate representation of the community sentiment in the 1940s, arguably turned out to be overblown. We will return to the foundational significance of the Schwinger-Tomonaga equation below, but its role in the empirical success of Dysonian perturbation theory was, at best, extremely thin.

Putting aside the issue of covariance for the moment, the more practical motivation for the interaction picture was its role in setting up a series expansion for the S-matrix in powers of H_I. Adopting a one-parameter family of space-like Cauchy surfaces, Dyson (1949b) equates the interaction picture time evolution operator with an iterative integration of the Schwinger-Tomonaga equation:

$$U(t_f, t_i) = \left(1 - i\int_{\tau_1}^{t_f} H_I(\tau)d\tau\right)\left(1 - i\int_{\tau_2}^{\tau_1} H_I(\tau)d\tau\right)\cdots, \tag{5}$$

where $\tau_1, \tau_2, ...$ are the time labels of the intermediate surfaces linking the final surface t_f to an initial surface t_i. Performing such a multiplication is easy when the Hamiltonian is

time-independent, as in the Schrodinger picture. One can simply take the Hamiltonian H out of the integrals and combine them into a single integral over the entire interval:

$$U(t_f, t_i) = \exp\left\{-iH \int_{t_i}^{t_f} d\tau\right\} = e^{-iH(t_f - t_i)}. \tag{6}$$

If, however, the Hamiltonian is time-dependent, as H_I always is in the interaction picture, one must take care to preserve the order of the operators when combining the integrals. What we end up with instead, therefore, is a "time-ordered exponential" series, commonly known as the Dyson series:

$$U(t_f, t_i) = T_\theta\left(\exp\left\{-i \int_{t_i}^{t_f} H_I(\tau) d\tau\right\}\right) = \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int_{t_i}^{t_f} dt_1 ... \int_{t_i}^{t_f} dt_n T_\theta[H_I(t_1)...H_I(t_n)]. \tag{7}$$

Here $T_\theta[...]$, is a compact notation for the expression:

$$T_\theta[H_I(t_1)...H_I(t_n)] = \sum_{p \in P_n} H_I(t_{p_1}) H_I(t_{p_2})...H_I(t_{p_n}) \theta(t_{p_1} - t_{p_2}) \theta(t_{p_2} - t_{p_3})...\theta(t_{p_{n-1}} - t_{p_n}) \tag{8}$$

where the sum is over all permutations p of n variables, and θ is the Heaviside step function.

These heaviside functions originate in the integration boundaries in equation (5). Jointly they ensure that operators which take earlier time arguments will always be to the right of those which take later time arguments, thus $T_\theta[...]$ is referred to as the time-ordered product (more on the unconventional θ subscript shortly). This time-ordering of operators in the expansion coefficients was interpreted by the architects of the causal perturbation theory program as a manifestation of another causality property, distinct from microcausality. Stueckelberg's original intuition (discussed in detail in section 4) was that a Hamiltonian operator associated with an earlier time must act on the quantum state first on pains of introducing causal connections from the future to the past into one's theory. Whereas the microcausality condition rules out superluminal influence (and thus concerns the behaviour of physical quantities outside of the lightcone), the causality condition which Stueckelberg and Bogoliubov saught was designed to rule out retrocausality (and thus concerns their behaviour inside the lightcone). Crucially, while in Dysonian perturbation theory the time-ordering of the expansion coefficients is a derived property, following automatically from the integration of the evolution equation, the basic idea behind causal perturbation theory was to start from this causal ordering property and use it to derive the perturbative expansion. This alternative derivation of perturbative QFT ultimately leads to an alternative definition of the time-ordered product which is not implemented via Heaviside functions; we have thus introduced a θ subscript to $T_\theta[...]$ above to distinguish it from the

time-ordered product $T_B[...]$ (after Bogoliubov) which one finds in the mature version of causal perturbation theory.

Using the relation,

$$S = U(\infty, -\infty), \tag{9}$$

Dyson obtained an expression for Heisenberg's S-matrix as a power series in H_I. Since H_I is a product of free field operators in the interaction picture, if we assume free particle in and out states each term in this expansion can be calculated, at least in principle. Indeed, Dyson popularised the use of Feyman diagrams to conveniently carry out such calculations. This improved formulation of perturbative QFT did not automatically solve the ultraviolet divergence problem, however. Starting at second order, the integrals over time-ordered products appearing in the series diverge. In the late 1940s, renormalization techniques were developed to handle these divergences, which Dyson integrated into his formalism. The basic idea behind the new renormalization procedure was to introduce a set of so-called counterterms to the theories Hamiltonian which had the effect of subtracting the divergent part of the series coefficients. This meant, in effect, introducing a new set of (finite) renormalized masses and interaction couplings, with the original (infinite) "bare" masses and couplings being cancelled by the (also infinite) counterterms. Once this procedure was carried out, truncations of Dyson's expansion for the QED S-matrix yielded stupendously accurate empirical predictions.

Despite this breakthrough, as we move into the 1950s the community assessment of Dysonian perturbation theory became increasingly negative. One reason for this rapid change of fortune was empirical: the startling new phenomena being produced in particle accelerators in the 1950s did not seem to be co-operating with the assumptions of the perturbative approximation scheme. Attempts to model nuclear interactions along the lines of perturbative QED quickly stalled, and it was felt that taming the strong interaction would require a radically different calculational approach than perturbation theory, predicated as it is on the weak strength of the interaction. More significantly for our story, the 1950s saw a new round of theoretical wrangling over the consistency of QFT.[4] Had Dysonian perturbation theory really resolved the challenges associated with the unification of quantum theory and special relativity? As the reception of the new formalism progressed the more foundationally orientated wing of the community came to feel that the problems facing Heisenberg-Pauli QED had been transmutted rather than conclusively solved by the new formalism.

The invention of perturbative renormalization did not put a stop to worries about the

[4]See Blum (2023) for a systematic discussion of the broader debate about the consistency of QED and QFT in the 1950s and 1960s. Here we only touch on a few aspects these developments which are relevant for the development of causal perturbation theory.

ultraviolet behaviour of QFT; if anything they intensified in the 1950s. One issue was the dubious mathematical rigour of the renormalization procedure. In Dyson's derivation of the expansion one had to write down and manipulate divergent integrals before replacing them with finite expressions. This had an ad hoc character, but also seemed to hinge on the manipulation of ill-defined quantities. The unrenormalized "bare" parameters, in particular, were identified with series containing divergent coefficients. It was thus unclear whether perturbative renormalization had really done anything to address the ultraviolet problem, or simply swept it under the rug (in a mathematically illegitimate way). The appearance of new, non-perturbative, arguments for the break down of QED on very short length scales—most famously the Landau pole problem (Landau et al. 1956)—seemed to favour the later interpretation.

These concerns about the ultraviolet break down of QFT also challenged the cogency of Dyson's derivation, since positing a differential evolution equation necessitated multiplying field operators at the same space-time point. While the Schwinger-Tomonaga equation pointed to the possibility of a fully covariant Hamiltonian dynamics for QFT, it was unclear whether Dyson's formalism delivered on this promise. Dyson himself had initially been optimistic that the perturbative expansion might be used to establish the existence of solutions of the field equations.[5] He ended up being amongst the first to conclude that, even after each term had been rendered finite via renormalization, the expansion as a whole did not converge (Dyson 1952): it turned out that renormalized QED perturbation theory was at best an asymptotic expansion, leaving the question of existence of solutions of the Schwinger-Tomonaga equation open. The continuing ultraviolet problems with QED suggested a negative answer. Furthermore, as we shall discuss further in section 5, Haag's theorem and the issue of boundary divergences suggested that the interaction picture evolution operator $U(t_f, t_i)$ did not in fact exist. While Dyson had undoubtedly provided an efficient algorithm for calculating the asymptotic S-matrix in the $t_f \to \infty, t_i \to -\infty$ limit, there was reason to question the coherence of his assumptions about finite time dynamics.

This all led to the return of Heisenberg's original project of seeking a formulation of the dynamics of relativistic quantum theories which is not based on a differential evolution equation; and this is where causality enters the picture. We have seen that causality properties naturally arise within Dyson's derivation of the perturbative expansion—microcausality, as a kind of necessary condition for making the switch to a fully covariant Schwinger-Tomonaga equation, and time-ordering of operators as a derived property which follows automatically from the formal integration of the differential evolution equation. The key idea of causal perturbation theory is that it is possible to derive the perturbative expan-

[5]See Blum (2023) chapter 2 for a detailed discussion of Dyson's thinking during this period.

sion of the S-matrix by starting from a time-ordering causality property. As we shall see, this alternative derivation would eventual lead to both a solution of some of Dysonian perturbation theories foundational problems and a clarification of its limitations.

4 Stueckelberg's Causality Condition

To tell the full story of the origins of causal perturbation theory we need to backtrack a little, as this alternative derivation of the perturbative expansion for the S-matrix originates in work by Ernst Stueckelberg which in fact preceeds that of Feynman, Schwinger, Tomonaga and Dyson. While Bogoliubov would eventually repackage the causal approach as a rigorous reconstruction of Dysonian perturbation theory, Stueckelberg conceived it as a novel solution to the problems facing relativistic quantum theory in the 1940s. Indeed, Stueckelberg explicitly styled his approach as a development of Heisenberg's S-matrix program.[6] He added two new elements to Heisenberg's original picture, however: firstly, he focused on the construction of a perturbative approximation (which Heisenberg had hoped to move beyond) and secondly, he introduced a causality condition as a further non-redundant constraint on the S-matrix. Stueckelberg's key insight in the early 1940s was that one could construct a unitary, Lorentz-invariant S-matrix that would be, in some presumed problematic sense, retrocausal. As we will see, the formulation of an appropriate causality condition evolved as the causal perturbation theory approach progressed, but let us start with Stueckelberg's original intuition.

In Heisenberg's scheme, the unitarity and Lorentz invariance of the S-matrix was ensured by writing it as the complex exponential of some hermitian Lorentz scalar η. An obvious ansatz suggested by Heisenberg is to identify η with the time integral of the interaction Hamiltonian $\int dt H_I(t)$. This is a hermitian Lorentz scalar and thus fulfils all of Heisenberg's criteria. However, when expanding this S-Matrix perturbatively to second order, one obtains:

$$e^{i\eta} \approx 1 - i\eta - \frac{1}{2}\eta^2 = 1 - \int_{-\infty}^{\infty} dt H_I(t) - \frac{1}{2}\int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} dt' H_I(t) H_I(t') \tag{10}$$

The second-order term can be split in two, one term for $t > t'$, another for $t < t'$, i.e.:

$$\int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} dt' H_I(t) H_I(t') = \int_{-\infty}^{\infty} dt \int_{-\infty}^{t} dt' H_I(t) H_I(t') + \int_{-\infty}^{\infty} dt \int_{-\infty}^{t} dt' H_I(t') H_I(t) \tag{11}$$

We thus get a term where the operator that corresponds to the later time acts first on the initial state. Expressing H_I in terms of creation and annihilation operators, this would

[6] A more detailed presentation of Stueckelberg's early work on the causal perturbation theory approach, and its relationship to Heisenberg's S-matrix program, can be found in Blum (2017).

correspond to the creation of a particle at a later time t and its annihilation at an earlier time t'. It was this circumstance that Stueckelberg identified as a violation of causality (Stueckelberg 1944). One might justifiably interject at this point that this analysis rests on a questionable reading of the physical content of the perturbation series. After all, we are talking about the time ordering of virtual events (if that term is even appropriate) in a single term of an expansion which really need not (and perhaps cannot) be interpreted separately from the series to which it belongs.

Putting aside the question of motivation for the moment, the central problem for Stueckelberg became how to systematically identify and eliminate these putatively problematic acausal contributions to the perturbation series. After some searching, Stueckelberg gave a first formulation of the causality condition in a short note in the Physical Review (Rivier and Stuecklberg 1948). When written explicitly as a volume integral over some product of field operators, all the terms in the perturbation expansion will contain some field operators that annihilate the particles in the initial state or create the particles in the final state with all other field operators being pairwise contracted, giving singular two-point functions—the propagators. Stueckelberg's causality condition amounted to the demand that all singular two-point functions that actually appear in the perturbation expansion of the S-matrix take the form of the "causal propagator" (also known as the "Feynman propagator") commonly written:

$$D_c(x,y) = \frac{i}{16\pi^3}\int \frac{d^3k}{\omega(\mathbf{k})} e^{-i\mathbf{k}(\mathbf{x}-\mathbf{y})}[\theta(x_0-y_0)e^{i\omega(\mathbf{k})(x_0-y_0)} + \theta(y_0-x_0)e^{-i\omega(\mathbf{k})(y_0-x_0)}] \quad (12)$$

Here $\omega(\mathbf{k})$ is the positive energy $\sqrt{m^2+\mathbf{k}^2}$ belonging to the momentum vector $\mathbf{k}$ and θ is the Heaviside function. D_c was read, by Stueckelberg, as describing the emission of plane waves of positive energy moving forwards in time and plane waves of negative energy moving backwards in time.

Given an interaction potential as an input, it turns out that this requirement fixes the form of the integrands that appear in the series coefficients and it was quickly shown that what one gets agrees with the first few terms of the Dyson series (Stueckelberg and Green 1951). On the face of it then, Stueckelberg's causal derivation of the perturbative expansion of the S-matrix was successful, and could, in principle, have provided an alternative basis for the development of renormalized QED in the post-war period. It was, however, largely ignored. The most important reason for this was likely bad timing; by the time Stueckelberg was capable of delivering concrete calculations, the famous second-order QED calculations of Schwinger and Feynman were widely known, and Dysonian perturbation theory had stolen his thunder. Stueckelberg's idiosyncratic notation and difficult-to-follow prose did not help matters.

Even putting these contingencies aside, however, it is fair to say that Stueckelberg's formalism had its fair share of deficiencies at this point. From a calculational perspective, his formalism compared unfavourably to the compact, user-friendly, Feynman rules. The conceptual foundations of this earliest version of causal perturbation theory were also rather murky. His causality condition was more of a recipe than a precise mathematical statement and, as we flagged above, its motivation hung on a dubious interpretation of the virtual processes represented by integrands appearing in the series coefficients. Furthermore, Stueckelberg's non-standard take on the problem of ultraviolet divergences, which we will discuss further in section 7, was not, at this early stage, well worked out.

If this had been the end of causal perturbation theory, it would rightfully have remained a curious footnote in the history of QFT. As it happened, however, the program did not peter out at this point but entered a new phase of development. In this second stage, the causal approach would be self-consciously styled as a way around some of the lingering problems with Dysonian perturbation theory, ultimately being conceived by Bogoluibov as a foundationally motivated reconstruction project rather than a novel theory. Bogoluibov introduced a clearer and better-motivated causality condition and worked out its consequences more systematically. Before we get to Bogoliubov's causality condition, however, we will first take a detour via another Stueckelbergian idea which turned out to have a crucial influence on its development: the appearance of another class of divergent integrals in the perturbative coefficients for finite (non-asymptotic) times.

5 Boundary Divergences

In September 1949, Stueckelberg attended the International Congress on Nuclear Physics and Quantum Electrodynamics, where Dyson first presented his formulation of perturbative QED to a European audience. From this point on the nature of Stueckelberg's rhetoric changes. Now that the basic goal of constructing predictively powerful perturbative approximations of scattering amplitudes had been achieved, he started to defend the superiority of his own framework along more foundational lines. Contrasting his own S-matrix first "integral" method with Dyson's differential approach, he asserted:

> The procedure of the integral method differs essentially from the differential method employed by Tomonaga, Schwinger, Feynman and Dyson in that the two sorts of diverging terms occurring in the formal solution of a Schrödinger equation are avoided. These two divergences are: 1) the well-known *"self-energy" divergencies* [sic] which have been since corrected by methods of regularization (Rivier, Pauli and Villars);[7] 2) the more serious *boundary divergen-*

[7] On the connection between the work of Stueckelberg's student Dominique Rivier and the more well-

> *cies* (Stueckelberg) due to the sharp spatio-temporal of the space-time region of evolution V in which the collisions occur. (Stueckelberg and Green 1951, 153)

The first type of divergence Stueckelberg gestures at here are the familiar ultraviolet divergences—we discuss how he proposed to address these within the causal perturbation theory approach in section 7. The second type of divergence, however, is likely unfamiliar even to QFT aficionados. We will discuss these "boundary divergences" in some detail here as in addition to posing yet another largely forgotten foundational problem with perturbative QFT, their analysis actually led (in a rather convoluted way) to Bogoluibov's improved causality condition.

Stueckelberg announces his discovery of these new divergences in a paper submitted to the Physical Review in the summer of 1950 (Stueckelberg 1951).[8] In Dyson's derivation, as we saw, the S-Matrix appeared as the infinite-time limit of the interaction picture time evolution operator $U(t_f, t_i)$, which was taken to describe the dynamical evolution linking the asymptotic scattering states. Stueckelberg had, in his earlier work, also toyed with the idea that his causal approach might also be applied to finite-time evolution by limiting time integrations to the region between an initial time t_i and a final time t_f. But no one, until now, had explicitly calculated such a finite-time evolution operator within perturbation theory. In his paper, Stueckelberg now argued that if one actually used Dyson's time evolution operator to "evaluate transition probabilities for processes which are localized in space-time by a *sharply defined boundary* (for example two time-like surfaces specifying an initial and final observation), one obtains divergent results." The paper is short and elliptical, but we will attempt, to reconstruct how Stueckelberg most likely arrived at this conclusion.

Stueckelberg approached the description of finite-time scattering in the following way. In the absence of interactions, the interaction representation wave function is constant. One could thus also obtain the finite-time evolution operator $U(t_f, t_i)$ by calculating the S-Matrix of a modified theory in which the interaction is "switched on" at time t_i and "switched off" at time t_f, i.e., where the interaction Hamiltonian is multiplied by a box function $g(t)$, which is equal to 1 when $t_f > t > t_i$ and zero otherwise. One could then take over a usual expression for the asymptotic S-Matrix (be it causal or Dysonian) and simply modify the form of the interaction term in order to describe finite-time dynamics.

Following this approach, Stueckelberg calculated the S-Matrix element for the propagation

known Pauli-Villars regularization method, cf. Schweber, QED.

[8] It seems plausible that Stueckelberg chose to publish this paper in the Physical Review – rather than in his usual journal, the Helvetica Physica Acta – because he conceived it as a direct response to Dyson's formulation of perturbative QFT.

of a single electron interacting with the electromagnetic field to second order in perturbation theory, with the QED interaction Hamiltonian being multiplied by the switching function $g(t)$:

$$\langle k_f|S(g)|k_i\rangle = \delta^{(3)}(\mathbf{k}_f - \mathbf{k}_i) + \int dx \int dy g(x_0)g(y_0)e^{-ik_f y}\overline{u}(\mathbf{k}_f)\Sigma(x-y)u(\mathbf{k}_i)e^{ik_i x}, \quad (13)$$

where k_i and k_f are the initial and final four-momentum of the electron, respectively, $\mathbf{k}_i$ and $\mathbf{k}_f$ the initial and final three-momentum, and u are the corresponding wave functions (all spin indices have been suppressed and the normalization conventions of Weinberg (1995) have been adopted). $\Sigma(x-y)$ is the self-energy of the electron at second order in perturbation theory, resulting from the emission and re-absorption of a virtual photon.

Transitioning to momentum space and introducing the Fourier transform of the switching function, $\tilde{g}(\omega)$, this becomes

$$\langle k_f|S[g]|k_i\rangle = \delta^{(3)}(\mathbf{k}_f - \mathbf{k}_i)\left[1 + \int d\omega \overline{u}(\mathbf{k}_f)\Sigma(\omega, \mathbf{k}_i)u(\mathbf{k}_i)\tilde{g}(\omega_f - \omega)\tilde{g}(\omega - \omega_i)\right]. \quad (14)$$

Now, in standard calculations of the asymptotic S-matrix, one sets $g(t) = 1$, and thus $\tilde{g}(\omega) = \delta(\omega)$, effectively imposing the on-shell condition, so that the self-energy contribution exactly cancels with the mass counterterm and the probability of an electron continuing with final momentum $k_f = k_i$ is simply 1. The introduction of a time-dependent switching function, however, leads to an integral over off-shell momenta. After a change of integration variables, one finds that the integral in equation (14) will converge only if the following expression converges:

$$\int d\omega \Sigma(\omega + \omega_i, \mathbf{k}_i)|\tilde{g}(\omega)|^2. \quad (15)$$

For the simple, instantaneous switching on and off of the interaction at times t_i and t_f respectively, where $g(x)$ is a box function, one has

$$|\tilde{g}(\omega)|^2 = \frac{4\sin^2\left[\omega\frac{(t_f - t_i)}{2}\right]}{\omega^2}. \quad (16)$$

Therefore, to obtain a convergent expression one would need $\Sigma(\omega, \mathbf{k}_i)$ to approach a constant value as $\omega \to \infty$ (while $\mathbf{k}_i$ remains fixed). Within perturbation theory, however, it grows linearly with ω leading to a divergent expression for the matrix element.[9]

[9]If one takes the renormalized expression for Σ at second order in perturbation theory from, e.g.,

What is to be made of this result? From a conceptual, and indeed purely formal, point of view this issue remains undertheorized to this day. Scattered references to Stueckelberg's boundary divergences do exist in later literature, but attitudes concerning the seriousness of the problem and how it should be resolved are quite varied.[10] There is a potential connection with Haag's theorem here. Haag (1955) put forward a non-perturbative argument for the impossibility of relating the states of a free and interacting QFT via a unitary transformation, indicating that Dyson's $U(t_f, t_i)$ interaction picture operator could not in fact exist;[11] Stueckelberg's boundary divergences likewise indicated a problem with $U(t_f, t_i)$, or at least its perturbative expansion. Note, however, that while Haag's theorem is often connected to infrared perturbative divergences, these boundary effects, in fact, lead to momentum space integrals which blow up in the region of arbitrarily large momentum—they are thus ultraviolet divergences, of a novel sort. Crucially, though, the conventional renormalization procedure which cures the usual ultraviolet divergences does nothing to alleviate these new infinities. Bogoluibov would later point out that even in the presence of a Pauli-Villars regulator divergences occur if one sharply switches on and off an interaction in some finite space-time region (Bogoliubov and Shirkov 1959).[12] Whatever the exact connection to Haag's theorem may be, the conclusion that Stueckelberg drew from the boundary divergences was that Dyson's $U(t_f, t_i)$ operator was ill-defined and thus his derivation of the perturbative expansion starting from the interaction picture evolution equation was mathematically faulty.

Stueckelberg thus used the boundary divergences as a new argument for the superiority of his causality condition-based derivation of the perturbative expansion (Stueckelberg and Green 1951). The new divergences could only be eliminated, he argued, by switching the

---

Weinberg (2002) Eq. 11.4.14, and takes the limit of large ω (with $\mathbf{k}$ fixed), one gets,

$$\Sigma(\omega) = -\frac{2i\pi e^2}{(2\pi)^4}\omega\gamma_0 \int_0^1 dx \left\{(1-x)\ln\left(\frac{m_e^2}{\omega^2 x}\right) - \left[(1-x)\ln\left(\frac{1-x}{x^2}\right) - \frac{2(1-x^2)}{x}\right]\right\}, \tag{17}$$

which, when plugged into equation (14) yields a non-convergent integral.

[10] See, for instance, Fredenhagen and Lindner (2014) and Baacke et al. (2001).

[11] As it happens, Bogoluibov (1951) contains statements which seem to anticipate Haag's theorem, though it is unclear how Bogoliubov reached these conclusions.

[12] There is a puzzling discrepancy between Bogoliubov's and Stueckelberg's discussions of boundary divergences. The Bogoluibov school refer only to the existence of boundary divergences in the self-energy of a boson (Sukhanov 1963), while the Stueckelberg (1951) calculation, reconstructed above, concerns the electron self-energy. Furthermore, neither Bogoliubov and Shirkov or Sukhanov explicitly derive integrals containing boundary divergences, instead employing rather indirect arguments to the effect that divergences must occur in the limit where the switching on (and off) of the interaction becomes instantaneous. We have thus not been able to pinpoint the origin of this discrepancy, which in any case did not seem to have impacted the conclusions that Stueckelberg and Bogoliubov each drew from the existence of the boundary divergences.

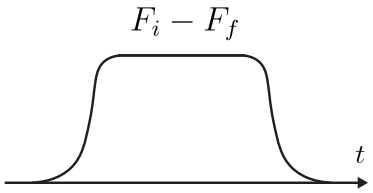

Fig 1

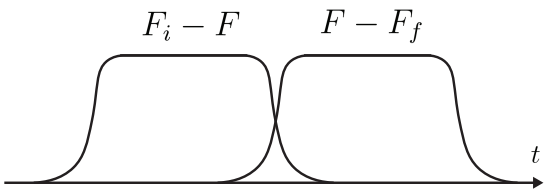

Fig 2

interaction on with a smooth rather than discontinuous function; this has the effect of suppressing the high-frequency components in $\tilde{g}$ leading to the convergence of the integral in equation (14). Finite time evolution was thus to be represented in the following way. One must first envision an initial state vector as depending not on a sharp time, but on a thin "time-like layer", represented by a smoothed out Heaviside function $F_i(t)$. Evolution to a later smoothed "layer", $F_f(t)$ could then be described by the generalized S-matrix:

$$\psi(F_f) = S(F_i - F_f)\psi(F_i), \tag{18}$$

$g(t) = F_i(t) - F_f(t)$ is now a smoothed-out box function (see figure 1) and the generalised S-matrix $S(g)$ is free from boundary divergences. Stueckelberg claimed that this new S-matrix could not be related back to a differential description of the time evolution; recovering a differential description would require taking the limit of $g(t)$ tending to a discontinuous function, spelling the return of boundary divergences. Stueckelberg thus contrasted a differential formulation of the dynamics of relativistic quantum theories with his own integral approach.

As before, Stueckelberg's latest offering received little attention in the West. It did find one receptive reader in the Soviet Union, however. In the late 1940s, Nicolay Bogoliubov was reading the latest Physical Review papers on renormalized QED with the mathematical-physics group at the Steklov Institute of Mathematics in Moscow (Medvedev 1994). Indeed, Bogoliubov's group appears to have been the first to engage with this subject in the Soviet Union (Kirzhnits 1994). Among the papers read was Stueckelberg's (1951) paper on boundary divergences, and he took it more seriously than most. It was Bogoliubov who would take causal perturbation theory to its next stage of development, formulating a precise causality condition and rebranding the program as a mathematical reconstruction project, as we shall see in the next section.

6 Bogoliubov's Causality Condition

In the fall of 1951, Bogoliubov (1952a,b,c) published a series of three short papers with the Russian Academy of Sciences which represent his first stab at the question of how the dynamics of a relativistic quantum theory ought to be formulated.[13] He embraced Stueckelberg's claim that, due to the boundary divergences, a smooth switching function would have to be built into the theory. Initially, he was somewhat critical of Stueckelberg's specific proposals about how this ought to be done, however.

First of all, he pointed out a puzzle with Stueckelberg's 'integral' representation of finite time evolution—equation (18). Suppose that we decompose the finite time evolution between the smoothed layers F_i and F_f into two parts; an initial period between F_i and some intermediate layer F, and a remaining period between F and F_f. Bogoliubov pointed out that applying Stueckelberg's prescription for generating a generalised S-matrix to the whole evolution and to the two sub-periods yielded different answers—that is $S(F - F_f)S(F_i - F) \neq S(F_i - F_f)$, due to the overlap between the smoothed box functions (see figure 2). Stueckelberg's brief discussion had thus left crucial questions about how his $S(g)$ matrices could be decomposed into products unanswered. Bogoliubov would ultimately realize that the factorization properties of the generalized S-matrix are intimately related to time-ordering causality, as we shall see.

Bogoliubov also questioned the claim that the adoption of a smooth switching function necessitated abandoning a differential evolution equation; indeed, his primary goal in this early trilogy of papers was to develop a generalization of the Schwinger-Tomonaga equation which incorporated the switching function. In order to achieve this Bogoliubov argued that, given certain assumptions, one could construct an analogue of the Hamiltonian corresponding to an $S(g)$ evolution operator. Now, as we saw in section 3, the functional dependence of the time-evolution operator and Hamiltonian is very complicated in the interaction picture. If the Hamiltonian were time independent, however, one could simply rearrange the relation $U = e^{-iHt}$ to obtain:

$$H = i\frac{\partial U}{\partial t}U^{\dagger}. \tag{19}$$

Bogoliubov (1952b) argued that this prescription actually worked more generally for time-dependent Hamiltonians, at least in the context of perturbation theory; if one inserts the Dyson series for $U(t_f, t_i)$ into the above equation, one gets the correct time-dependent interaction picture Hamiltonian due to the cancellation of higher-order terms in this ex-

[13] Translations of all three papers can be found in (Bogolubov Jr. 1995). We would like to thank Kseniia Mohelsky for providing translations of all other Russian-language sources used in this paper.

pansion.[14]

He therefore proposed defining a generalized Hamiltonian density $\mathcal{H}(x, g)$ corresponding to the evolution operator $S(g)$ (with the switching function $g(x)$ now understood as a function of space as well as time) as follows:

$$\mathcal{H}(x, g) = i\frac{\delta S(g)}{\delta g(x)} S^{\dagger}(g). \tag{20}$$

One could then conceive of a functional differential equation which describes the variation of the state under variations of the switching function $g(x)$:

$$i\frac{\delta \psi(g)}{\delta g(x)} = \mathcal{H}(x, g)\psi(g). \tag{21}$$

Taking $g(x)$ to be Heaviside functions centred on space-like hypersurfaces takes us back to the usual Schwinger-Tomonaga equation. This would be accompanied by boundary divergences, however, indicating that an evolution equation linking infinitely thin surfaces is not well-defined.[15] Nevertheless, in these early papers Bogoliubov seemed to suggest that one could still view the generalized functional equation (21), describing the advancement of smoothed 'layers' rather than Cauchy surfaces, as the core statement of a QFTs dynamics, contra Stueckelberg.

By the time Bogoliubov wrote on QFT again in 1955, however, his perspective had changed (Bogoliubov 1955). In the interim, he had delved deeper into Stueckelberg's early work and now explicitly advocated the priority of a causality condition over a Hamiltonian evolution equation, though for his own distinctive reasons.[16] He writes:

> [I]t is desirable to have a representation of quantum field theory that would allow us to see the basic physical assumptions on which it is built in its modern form, in order to be able to understand in which directions it is acceptable to generalize them. In the usual representation of the quantum field theory, based on the Hamiltonian formalism, those assumptions do not receive the proper attention. In our opinion, it is much more useful to proceed from

[14] Bogoliubov (1952b) claims that this cancellation of higher-order terms holds if one assumes a micro-causality property, representing the first appearance of causality considerations in his work on QFT.

[15] Since boundary divergences had only been shown to appear in the perturbative expansion coefficients one could question whether this problem also occurred in a non-perturbative formulation. Bogoliubov, in any case, consistently took the boundary divergences to problematize the conventional Schwinger-Tomonaga equation; see discussions in Bogoliubov and Shirkov (1959) chapter 6.

[16] It is likely that when Bogoliubov wrote his trilogy of papers on the dynamical equations of QFT he had only read Stueckelberg (1951), and was therefore unaware of Stueckelberg's broader causal perturbation theory project, which is not described in that short paper.

> the scheme suggested by Stueckelberg [...], where he introduces a generalized S-matrix without referring to the Hamiltonian formalism. The role of the Hamiltonian formalism in specifying the form of the S-matrix is taken by clearly formulated physical conditions, among which a causality condition is the basic one. (Bogoliubov 1955, pp. 237)

Notice that, unlike Stueckelberg, Bogoliubov does not claim that a Hamiltonian description of the dynamics is impossible. Rather, his mature view is that causality is more fundamental; while it may be possible to obtain an evolution equation like (21) this has a derived status, since in Bogoliubov's new causal perturbation theory formalism the $S(g)$ operator is constructed first using only his causality condition. Where Stueckelberg had conceived of causal perturbation theory as a rival formalism, Bogoliubov now presented his version of causal perturbation theory as a more conceptually and mathematically rigorous reconstruction of Dysonian perturbation theory.

It is instructive to compare Bogoliubov's incarnation of the causal perturbation theory program with the axiomatic QFT program pursued by figures like Arthur Wightman and Rudolf Haag in the same period. By the mid-1950s any initial optimism generated by the empirical success of renormalized QED had largely evaporated and worries about the consistency of QFT once again loomed large. In this context, a number of theorists advocated a clean-up operation, in which a higher standard of mathematical rigour would be brought to bear and the theory's basic assumptions would be made explicit. We can see Bogoliubov, Wightman and Haag as key figures in this broad trend. Furthermore, Bogoliubov's project was also 'axiomatic' in the sense that it aimed to identify a fundamental set of principles underlying an area of physics, and both causal perturbation theory and axiomatic QFT would both give pride of place to a causality condition in their reformulation efforts.

There was a crucial difference regarding their attitude to Dysonian perturbation theory, however. The axiomatizations of QFT developed by Wightman and Haag were explicitly non-perturbative in character, eschewing all appeals to a perturbative expansion. As we mentioned at the end of section 3, there were strong empirical and foundational motivations for trying to move beyond perturbation theory in the 1950s and 1960s, so this approach was not unreasonable. Bogoliubov's methodological stance was quite different on this point, however. He opted to start from the conventional perturbative formalism, which he claimed "best corresponds to the actual modern state of field theory, where up to the present various formal expansions in powers of the smallness of the interaction cannot be removed and where all fundamental results were obtained with the help of these expansions" (Bogoliubov 1955, pp. 237). To use a political analogy, Wightman and Haag were more revolutionary in their attempts to develop a new non-perturbative language

for relativistic quantum theory, whereas Bogoliubov played the role of a reformer, incrementally improving the rigour of the tried and tested perturbative approximation scheme. This meant that the goals of Bogoliubov's causal perturbation theory were more modest —causal perturbation theory could not resolve the non-perturbative foundational issues that were coming to the fore in the 1950s, for instance (we return to this theme in section 8).

This difference in orientation towards perturbation theory had important implications for the formalizations of relativistic causality which these programs adopted. Following LSZ, Wightman adopted microcausality into his system of axioms, requiring that space-like separated field operators commutators to vanish, and this was eventually incorporated into the later algebraic axiomatizations of QFT. As Stueckelberg had urged, however, it was really the time-ordering property of the S-matrix which mattered in the perturbative context. While Bogoliubov embraced the idea of using causality to construct the S-Matrix, he found Stueckelberg's attempts to formulate a causality condition for that purpose to not be "sufficiently clear and general", explaining why his "ideas did not receive wide recognition" (Bogoliubov 1955, 237). As we saw in section 4, Stueckelberg's condition was rather vague mathematically and never led to a systematic account of the structure of the expansion to all orders. One of Bogoliubov's key contributions, therefore, was to put forward a new formulation of the causality condition which could be imposed directly on the S-matrix.

In order to formulate his causality condition, Bogoliubov repurposed the smooth switching functions he had worked with in his earlier papers. Though originally introduced to tame the boundary divergences, these functions now took on a life of their own, providing a language for describing dependencies between localized events within a pure S-matrix formalism. This required some further generalizations of Stueckelberg's ideas. Rather than restricting oneself to smoothed box functions in the time variable, as Stueckelberg had done, Bogoliubov worked with arbitrary smooth functions of space-time, $g(x)$, taking values between 0 and 1; one could understand these functions as smoothly turning the interaction on-and-off in different regions of space-time.

Bogoliubov, in fact, put forward two versions of his new causality condition: a differential and integral form. The differential form seems to have come first historically, likely evolving from his earlier ideas about generalizing the Hamiltonian formalism; we briefly sketch its motivation here. Consider how making an infinitesimal change to the switching function at a point x will affect the outgoing state that results from the $S(g)$ operator. In order to enforce time-ordering causality, and rule out retrocausal influence, Bogoliubov reasoned that the infinitesimal change in the wave function caused by an infinitesimal change in the switching function at point x should be independent of the shape of the

switching function at points earlier than x. One can imagine the change in the switching function as a ripple that propagates forward in time to affect the final-state wave function, its propagation being affected by the future shape of the switching function, but not by its past. Mathematically, Bogoliubov expressed this requirement in the following form:[17]

$$\frac{\delta}{\delta g(y)}\left(\frac{\delta S(g)}{\delta g(x)}S(g)^{\dagger}\right)=0, \qquad x \gtrsim y \tag{22}$$

where the $\lesssim$ symbol indicates that the point y at which a variation in the switching function is being made is either in the past with respect to x or is space-like separated (i.e. it is not in the future light cone of x).

In his textbook with his student Dimitri Shirkov, Bogoliubov introduced an equivalent integral version of his causality condition which we shall use to sketch the causal derivation of the series expansion (Bogoliubov and Shirkov 1959). In our view, this version of the causality condition has a more intuitive physical interpretation and more directly connects to the perturbative time-ordering property discussed in section 3. It is also this version of the causality condition which has been adopted in contemporary mathematical physics developments of causal perturbation theory.[18] To motivate this version of the causality condition it is useful to return to the question Bogoliubov had originally raised about Stueckelberg's generalised $S(g)$ matrix: under what conditions can we decompose such an S-matrix into a product of operators associated with sub-sections of the full scattering process? Essentially, Bogoluibov's integral causality condition asserts that Stueckelberg's composition of two S-matrices $S(g_1)$ and $S(g_2)$ holds only when the two smoothed switching functions g_1 and g_2 do not have overlapping support. Suppose that g_1 and g_2 have support in two non-overlapping regions G_1 and G_2, where none of the points in G_1 are in the future light cone of any of the points in G_2, then the causality condition states that:

$$S(g_1+g_2)=S(g_2)S(g_1), \quad G_2 \gtrsim G_1 \tag{23}$$

This can be read as stating that the effect of an earlier period of scattering, in region G_1, on the outgoing states is independent of that of a later period of scattering, in G_2, ruling out retrocausal influence. By the same token, it requires that, if the S-matrix can be decomposed into a product it takes the form of a time-ordered product, with the operator associated with the earlier period of scattering acting on the state first.

[17] Notice that, using the generalized Hamiltonian which Bogoliubov had introduced in his trilogy this condition can be written $\frac{\delta}{\delta g(y)}H(x,g)=0$, for $y \lesssim x$. This supports the conjecture that Bogoliubov's earlier tampering with a generalized Schwinger-Tomonaga equation led directly to the differential formulation of his causality condition.

[18] As one sees in Epstein and Glaser (1973).

One advantage of Bogoliubov's new causality condition (in both its differential and integral form) is that it does not rest on a physical interpretation of individual terms in the series expansion as Stueckelberg's had.[19] Indeed, thus far these conditions have been formulated as fully non-perturbative properties. Bogoliubov's other key advance over Stueckelberg's formalism was to demonstrate that one could derive the form of the Dyson series to all orders from his causality condition. Since this is really the key result of causal perturbation theory we sketch a version of this derivation here.[20]

With the introduction of the switching function, the general form of an expansion for the S-matrix can be written:

$$S(g) = \sum_{n=0} \frac{1}{n!} \int S_n(x_1, ..., x_n) g(x_1) ... g(x_n) dx_1 ... dx_n, \tag{24}$$

where $S_n(x_1, ..., x_n)$ are required to be Lorentz scalars but are otherwise left unspecified ($S_0 = 1$). If we plug this expansion into either side of Bogoluibov's integral causality condition and rearrange the resulting terms one can obtain two equivalent series:[21]

$$\begin{aligned} S(g_1 + g_2) = & \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{m!(n-m)!} \int d^4x_1 ... d^4x_n S_n(x_1, ..., x_n) \\ & \times g_2(x_1) ... g_2(x_m) g_1(x_{m+1}) ... g_1(x_n) \\ S(g_1) S(g_2) = & \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{m!(n-m)!} \int d^4x_1 ... d^4x_n S_m(x_1, ..., x_m) S_{n-m}(x_{m+1}, ..., x_n) \\ & \times g_2(x_1) ... g_2(x_m) g_1(x_{m+1}) ... g_1(x_n). \end{aligned}$$

Where the time components of $\{x_1, ..., x_m\}$ are all later than $\{x_{m+1}, ..., x_n\}$ in some reference frame. It follows that,

$$S_n(x_1, ..., x_n) = S_m(x_1, ..., x_m) S_{n-m}(x_{m+1}, ..., x_n). \tag{25}$$

Using this relation we can determine higher-order terms in the series inductively from the lower-order terms. Making the identification $S_1(x) = i\mathcal{H}_I(x)$, which amounts to a choice

[19] As a curious aside, Stueckelberg and Petermann (1953) try to give a non-perturbative statement of causality; we have been unable to decipher the meaning of the complicated condition they write down, however, and they do not try to use it to justify Stueckelberg's perturbative causality condition.

[20] Our discussion here follows the modern presentation of Scharf (2014), but a similar derivation is found in Bogoliubov and Shirkov (1959).

[21] The derivation given here follows the presentation of Scharf (2014). Bogoluibov proceeded by first obtaining a differential version of his causality condition.

of interaction term for one's perturbative model, one obtains higher-order terms of the form:

$$S_n(x_1, ..., x_n) = (i)^n T_B[\mathcal{H}_I(x_1)...\mathcal{H}_I(x_n)], \tag{26}$$

where $T_B[...]$ indicates that the product of Hamiltonian operators is time-ordered (more on the B subscript shortly). As in Dyson's derivation, therefore, what one ends up with is a series consisting of time-ordered products of the interaction Hamiltonian. If one takes the limit $g(x) \to 1$, returning to the usual asymptotic S-matrix operator, one indeed obtains the standard expression for the Dyson series. Bogoliubov was thus able to demonstrate that starting from the assumptions of Lorentz invariance, unitarity and his causality condition, one could derive the usual perturbative expansion of the S-matrix without making any use of a Hamiltonian evolution equation.

7 Causality and Renormalization

Bogoliubov's reformulation project was not intended to furnish new predictions or more efficient techniques for computing series coefficients; the goal was to illuminate the conceptual structure of perturbative QFT. One immediate conceptual result established by what we have seen thus far is that Dyson's expansion for the S-matrix can be obtained from a more minimal set of assumptions. Where time-ordering causality followed from the integration of the Schwinger-Tomonaga equation in Dyson's derivation, Bogoliubov showed that one could treat this property as primitive and avoid making any appeal to an evolution equation. This was significant because, by the mid-1950s, the status of the Schwinger-Tomonaga equation was in question. While the boundary divergences which concerned Stueckelberg and Bogoliubov were never widely discussed, non-perturbative problems, such as the Landau pole, were taken more seriously, and pointed to the potential non-existence of solutions of such an equation. We can see Bogoliubov's work on causal perturbation theory as attempting to disentangle perturbation theory from these non-perturbative problems, allowing the internal issues with the perturbative formalism to be addressed separately. Ultimately, this modest strategy would lead to a major victory; while these results remain surprisingly little known today, causal perturbation theory would form the basis of a mathematically rigorous treatment of the perturbative ultraviolet divergences problem, as we shall discuss in this section.

Recall that in Dysonian perturbation theory the ultraviolet divergences appearing in the expansion coefficients are 'subtracted' via the introduction of counterterms. In order for this to work the counterterm parameters and the original so-called 'bare' masses and coupling constants have to be equated with divergent expressions. While perturbative renormalization was certainly empirically successful it seemed to many to be both ad hoc

and mathematically dubious. There was still a latent sense that the ultraviolet divergences in QED perturbation theory pointed to the pathological short-distance behaviour of the theory, an impression which was bolstered by the Landau pole problem.

From the beginning in Stueckelberg's early work, causal perturbation theory had been tied to a very different approach to the ultraviolet divergence problem. Staying in position space, ultraviolet divergence can be viewed as stemming from the integration of products of propagators over short distances, or more precisely over the points at which their space-time arguments coincide. Intuitively, since considerations of causality concern the temporal ordering of events they tell us nothing about how quantities in one's theory ought to behave at these coincident points. This led Stueckelberg and Rivier (1950), in their first systematic presentation of the causal perturbation theory approach, to state that the causal propagator ought to be regarded as ambiguously defined at the origin. Bogoluibov would likewise argue that his causality condition does not uniquely fix the form of the series coefficients. Returning to the derivation of the series expansion given in the previous section, Bogoliubov pointed out that (25) is not actually the most general form for $S_n(x_1, .., x_n)$ which satisfies his causality condition. Starting at second order, we can add what Bogoluibov called 'quasi-local operators', Λ_n, to each term: products of Dirac delta functions $\delta(x_1 - x_2), ..., \delta(x_1 - x_n)$ and their derivatives (Bogoliubov and Shirkov 1959, chapter 4); since these operators only modify the behaviour of $S_n(x_1, ..., x_n)$ at $x_1 = ... = x_n$ there is no conflict with causality. Deriving the series expansion from a causality condition already suggested a different interpretation of the issue with the higher-order coefficients of expansion then; the novel feature of these terms was not so much that they contained infinities but that they contained ambiguities.

Why then do ultraviolet divergent integrals appear in the conventional Dysonian treatment? Answering this question, and developing this alternative ambiguity-based interpretation of perturbative renormalization, would require the importation of new mathematical concepts into the causal perturbation theory program. As it happened, the necessary resources had just been developed in pure mathematics, with Laurent Schwartz's influential books on distribution theory appearing in 1950 and 1951.[22] Distributions are objects which generalize the standard notion of a function. It is possible to uniquely associate a locally integrable function, $f(x)$, with a functional that takes test functions, $g(x)$, to the numbers:

$$T_f : g(x) \rightarrow \int_{-\infty}^{\infty} f(x)g(x)dx; \tag{27}$$

The basic idea of distribution theory is to consider a larger class of functionals that includes

[22]Schwartz's work on distributions in fact ran neatly parallel with the development of renormalized perturbation theory—see Barany et al. (2017) for a historical account. Note that Bogoliubov was likely also drawing on the older work of Sobolev—see footnote 23.

more singular objects, such as the Dirac delta 'function',

$$T_\delta : g(x) \to \int_{-\infty}^{\infty} \delta(x)g(x)dx = g(0), \tag{28}$$

the archetypal example of a singular distribution. Crucially, whereas operations like differentiation and the Fourier transform generalize to this larger set of objects, pointwise multiplication does not. Indeed, the product of singular distributions is not generally well-defined. Note that using $g(x)$ to represent the test functions is deliberately suggestive notation, as the switching functions would, once again, be repurposed to act as test functions in causal perturbation theory.

When it came to relating these mathematical innovations to perturbative QFT the key realization was that the time-ordered products appearing in the expansions coefficient are in fact products of singular distributions and therefore stand in need of an additional definition. Somewhat remarkably, Bogoliubov and Stueckelberg seem to have made this connection independently.[23] Bogoliubov already wrote in 1952 that the perturbative coefficients contained "product of singular functions" which required a "special definition", the absence of which being the cause of the "ultraviolet catastrophe" (Bogoliubov 1952c). A year later, Stueckelberg, in the final and most ambitious incarnation of his version of causal perturbation theory, explicitly integrated Schwartz's notions into his analysis of the series coefficients, writing:

> Unlike recent formalisms (Dyson and others) in which the divergences are accepted as such and "renormalized" by means of an algebra of infinite quantities [...] we consider that the multiplicative products of distributions T of $A, B...$ that is to say $T = AB...$ are in general not defined. (Stueckelberg and Petermann 1953, 509)

For both Stueckelberg and Bogoliubov, the causal derivation of the expansion was seen as a better starting point for providing a mathematically precise definition for the products of distributions appearing in the coefficients than the conventional Dysonian derivation.

[23] Bogoliubov (1952) already introduces the connection between renormalization and distributions, though he does not use that term. It is likely that Bogoliubov was drawing on knowledge of the Russian mathematician Sobolev's earlier concept of generalized functions which predated and influenced Schwartz's theory of distributions. Stueckelberg seems to have encountered distribution theory through interactions with Georges de Rham, a prominent Swiss mathematician who, like Stueckelberg, held positions at the universities of Lausanne and Geneva. De Rham was also an early adopter of Schwartz's ideas about distributions, and apparently recommended one of his mathematics students, André Petermann, to work with Stueckelberg in applying these concepts to QFT, leading to the Stueckelberg and Petermann (1953) paper (thanks to Gérard Wanders for conveying these detail to us).

Connections between distribution theory and QFT seem to have been in the air in the 1950s. Wightman's non-perturbative axiomatization of QFT, in particular, would also make use of the new notions, treating quantum fields as operator-valued distributions.[24] Note, however, that recognizing that the perturbative coefficients contain products of distributions does not require one to buy into Wightman's formalism, or indeed to causal perturbation theory. In fact, one can see by looking at the textbook formula for the causal/Feynman propagator that it is a distribution with a singularity at $x - y = 0$.[25] This follows simply from the fact that,

$$D_c(x,y) = \frac{i}{16\pi^3}\int \frac{d^3k}{\omega(\mathbf{k})} e^{-i\mathbf{k}(\mathbf{x}-\mathbf{y})}[\theta(x_0-y_0)e^{i\omega(\mathbf{k})(x_0-y_0)} + \theta(y_0-x_0)e^{-i\omega(\mathbf{k})(y_0-x_0)}] \quad (29)$$

contains Heaviside step 'functions', which are themselves singular distributions. As we highlighted in section 3, the Heaviside functions in the conventional time-ordered product, $T_\theta[...]$, arise automatically from the formal integration of the Schwinger-Tomonaga equation performed by Dyson. One can view the divergent integrals in the coefficients as arising due to this implementation of the time-ordered product, which then needs to be corrected by an infinite subtraction procedure. By treating time-ordering causality as an axiom rather than a derived property, however, Bogoliubov was able to treat the definition of the time-ordered products more carefully. This is why we used the notation $T_B[...]$ in the previous section to highlight that Bogoliubov's derivation of the expansion does not immediately lead to products of Heaviside functions; rather, Bogoliubov's characterization of the time-ordering property leaves the behaviour of the product at coincident points unspecified.

With Bogoliubov's rebranding of causal perturbation theory as a mathematical reconstruction project, the goal was to show that one could carry out a distribution theoretic construction of the products appearing in perturbation theory, reproducing the results of conventional renormalization without invoking an infinite subtraction. Bogoliubov suggested the following construction procedure:

> First of all, we need to define the indicated functionals for the special class of test functions which, together with all derivatives to some order, go to zero if any two points $x_1, \ldots, x_n$ match. After that, we need to extend those linear functionals to a class of arbitrary regular test functions. (Bogoliubov 1952b)

The causality condition fixes the behaviour of S_n everywhere except at coincident spacetime points; it thus allows us to construct a time-ordered product on a space of test

[24] See Wightman (1996) for a retrospective discussion.

[25] See Helling (2012) for an illuminating contemporary discussion of the application of distribution theory to products of QFT propagators.

functions (i.e. switching functions $g(x)$) which vanish at those points. This is perfectly well-defined since the switching functions vanish at the singularities. Renormalization was now mathematically recast as a problem of determining the extension of this product to the full space of test functions, i.e. to switching functions which are non-zero at coincident space-time points.

Bogoliubov and Stueckelberg both claimed that this extension exists but is not unique: the presence of products of singular distributions in the coefficients leads to delta function type ambiguities of the type left open by the causality condition.[26] In order to address the worry that these ambiguities render the resulting series meaningless or non-predictive Stueckelberg and Petermann (1953) introduced the notion of the renormalization group—a set of transformations between different ways of fixing the ambiguities which correspond to different, but empirically equivalent, definition of the expansion parameter. Furthermore, Bogoliubov pointed out that the ambiguities one gets from extending the distributional products correspond exactly to ambiguities which also appear in the conventional subtraction procedure, since when counterterms are added to cancel the divergences one also has to fix an arbitrary finite contribution (Bogoliubov and Parasiuk 1957). In modern parlance, this corresponds to the freedom to select different renormalization schemes. Thus, Fraser (2021) argues that causal perturbation theory's focus on renormalization ambiguities in fact had an important, but largely forgotten, influence on the development of key concepts like the renormalization group and the renormalization scheme which inform how perturbative QFT is understood in contemporary high energy theory.

While Stueckelberg and Bogoliubov set out a clear vision for a distribution theoretic reformulation of the perturbative renormalization procedure it is fair to say that they did not bring this project to fruition with a high standard of mathematical rigour. This was done later, however, by later mathematical physicists, most notably Epstein and Glaser (1973), who adopted the causal derivation of the series expansion and used it as the basis for a fully explicit distribution theoretic analysis of the higher-order coefficients of the expansion. While these results remain relatively unknown in mainstream high-energy physics, the causal perturbation theory tradition actually produced a mathematically rigorous resolution of the ultraviolet divergences problem. From the perspective of this later mathematically mature articulation of causal perturbation theory, ultraviolet divergences

[26] One can also construct perturbation series in quantum mechanics using a version of Bogoliubov's causality condition rather than the Schrodinger equation—see Scharf (2014). In this case, however, the perturbative coefficients are uniquely fixed by the information the causality condition provides about non-coincident points. It is the singular nature of the propagators in QFT which makes possible the addition of quasi-local operators which are genuinely unfixed by the causality condition. Furthermore, it is the strength of the singularity of the factors which determines the form of the ambiguity which arises in the product—see Helling (2012) for a discussion of this.

arise in the conventional approach due to a naive treatment of products of singular distributions appearing in the series expansion, which is forced upon one in the standard Dysonian derivation. Proceeding via the causal derivation it is possible to instead construct each term in the series without ever writing down or manipulating a divergent expression.

Somewhat ironically, Heisenberg's original intuition that adopting a more minimal dynamical framework would help resolve the ultraviolet divergence problem was vindicated, but in a completely different way from how he imagined. Heisenberg understood perturbative ultraviolet divergences to indicate a physical breakdown of QFT, necessitating the introduction of a fundamental length. The mathematization of the ultraviolet divergences problem found in contemporary developments of causal perturbation theory points to the opposite conclusion, however. From this perspective, ultraviolet divergences do not indicate the physical breakdown of QFT as short-length scales, they are rather unmasked as mathematical artefacts stemming from an incorrect treatment of the relevant distributional products. The interpretative implications of all this deserve more careful philosophical analysis than we can provide here, but it is clear that the treatment of renormalization found in causal perturbation theory makes a major contribution which philosophers working on the foundations of QFT need to engage with.

8 Causality and the Formulation of Relativistic Quantum Theory

Having now examined the causal perturbation theory program in some detail, we return in this section to the broader context of the 1950s and 1960s high energy theory and compare causal perturbation theory to other causality-based reformulation projects such as the axiomatic QFT and dispersion relations traditions. In section 2, we suggested that the influence of Heisenberg's S-matrix program in the 1950s was greater than has often been appreciated. One way to frame this continued relevance is to see many theorists in this period as attempting to plot a course between Heisenberg's original S-matrix theory and Dyson's formulation of perturbative QFT. It was clear that Heisenberg's principles of unitarity and Lorentz invariance were too minimal a starting point for relativistic quantum theory. They appeared to be too permissive; Stueckelberg was not the only author arguing already in the 1940s that models with a Lorentz invariant S-matrix could still be non-causal in some problematic sense (see the discussion of Kramers and Kronig and the dispersion relations tradition below). But they were also incapable of delivering concrete quantitative results; one could not conceivably formulate a method for calculating S-matrix elements starting from unitarity and Lorentz invariance alone.

Dysonian perturbation theory responded to both of these deficiencies, positing a much richer dynamical structure and using it to derive a powerful approximation scheme. In the course of the critical reception of this new formalism that took place in the following decades, however, it came to be felt that rather too much had been added to Heisenberg's original principles. The Schwinger-Tomonaga equation seemed to play an ephemeral, and purely formal, role in the derivation of the expansion and doubts about the possibility of constructing solutions to such an equation remained, and indeed intensified in the 1950s. Furthermore, it was increasingly suspected that the long-sought theory of the strong nuclear interaction would lie outside the scope of the Hamiltonian quantization scheme that Dyson worked to rehabilitate. As we stressed in section 3, the Schwinger-Tomonaga framework incorporated causal properties such as microcausality and perturbative time-ordering that did not follow from Heisenberg's principles of Lorentz invariance and unitarity. The question was whether these properties could somehow be abstracted from the more problematic aspects of the Dysonian framework and worked with directly. This would amount to articulating a middle ground between Heisenberg S-matrix theory and Dyson's perturbative formalism.

It should already be clear that causal perturbation theory (especially as developed by Bogoliubov) can be understood as a manifestation of this impulse. We want to highlight, however, that the methodological path taken by causal perturbation theory was a relatively modest and conservative one in comparison to other attempts to use causality properties to articulate a mid-point between Heisenberg and Dyson that flourished in this period. What Bogoliubov essentially did was start with the conventional perturbative approximation scheme and ask whether the same results could be derived from a weaker set of assumptions. In his hands, the causal perturbation theory approach became a reconstruction project rather than an exercise in novel theory building. Many theorists of the period were hoping for something much grander from a causality-condition-based reformulation of relativistic quantum theory, however. There was a dream that finding a stable middle ground between Heisenberg and Dyson would also lead to new non-perturbative calculational resources.

As we saw in section 7, the most important foundational upshot of the causal perturbation theory program was a new analysis of the ultraviolet divergences problem and the nature of perturbative renormalization. This was not as big news in the 1950s as it might have been in the 1930s, as the problem of ultraviolet divergences was now one foundational problem among many facing QFT. Causal perturbation theory did nothing to address the large-order behaviour of the Dyson series, for instance, it simply reproduced it; the series expansion obtained from the causal derivation was presumably also asymptotic. The distribution theoretic approach to perturbative ultraviolet divergences suggested that they do not in fact spell the doom of the theory, but it also left the Landau pole problem

untouched, suggesting that QED breaks down anyway for different (non-perturbative) reasons. Thus, while causal perturbation theory helped improve the internal coherence of the perturbative treatment of interactions, it was clear that it could not bring a resolution to all of the foundational issues plaguing QFT. The need to go beyond perturbation theory in order to finally slay QFT's demons was one of the key motivations for axiomatic QFT, which also put causality conditions centre stage but unlike Bogoliubov attempted to use them to formulate a new non-perturbative language for relativistic quantum theory.

While it might initially seem surprising to view Heisenberg's S-matrix theory and axiomatic QFT as allied programs, there are in fact clear continuities between them. The famous LSZ formalism of Lehmann et al. (1955), which represents a crucial origin point for the axiomatic tradition, follows Heisenberg in proposing a formulation of relativistic quantum theory based on globally imposed conditions. LSZ added back the field operators, which Heisenberg had hoped to eliminate, but they eschewed appeals to a local evolution equation. In fact, the main role played by the field operators in the LSZ formalism is to implement a principle of microcausality—the vanishing of the commutator of space-like separated field operators—which together with an asymptotic condition needed to establish the connection with the S-matrix was supposed to underwrite a formulation of relativistic scattering theory that does not invoke the perturbative expansion. As Arthur Wightman would say when discussing the importance of the LSZ formalism:

> Initially, I believe that LSZ took the view that it was a considerable advantage to work with their formalism because you didn't have to go down to the disgusting problems of Lagrangian field theory. To some extent when you have a new formulation of things you can celebrate that and contrast it with the old. (Wightman interview with Mehra)

Wightman goes onto say that the aspiration in the early days of axiomatic QFT was to "try and extract completely the content of the axiom as opposed to the content of specific dynamics" and thus avoid engaging with the problematic Hamiltonian based quantization proceedures which Dysonian perturbation theory remained tied to.

This optimism about the possibility of extracting non-perturbative information from causality properties was, we conjecture, bolstered by the emergence of dispersion relations as a potential calculational alternative to perturbation theory. Building on the work of Kramers and Konig in the 1920s, a connection was drawn in the 1950s between causality and the behaviour of S-matrix elements on the complex plane (see Cushing (1990) for a discussion of this early work). Dispersion relations—formula relating the real and imaginary part of scattering amplitudes—were seen as encoding causal structure, but by the same token as derivable from causality principles, thus presenting a new route from causal assumptions to non-perturbative quantitative results. Notably, Goldberger (1955) argued that the ana-

lyticity properties needed to derive dispersion relations followed from microcausality. The absence of singularities in the complex plane thus came to be seen as yet another way of imposing relativistic causality on one's theory. Chew's S-matrix program was the most ambitious implementation of this dispersion relations approach, proposing to derive the S-matrix of strongly interacting systems from assumptions about its analyticity properties. It is worth pointing out, however, that in this period there were substantial interactions between axiomatic QFT and this more phenomenologically orientated wing of high energy theory, with many of the pioneers of axiomatic QFT also working on the derivation of dispersion relations (see papers in Klein (1961)). These connections remain underexplored but do suggest that, while there are certainly important differences between early axiomatic QFT and Chew's bootstrap theory, we ought to see them as part of a more ambitious theory-building project based on causality properties.

In the end, the idea that causality-based reformulations of QFT would lead to a sweeping non-perturbative approach of the theory proved to be utopian. While both the axiomatic QFT and dispersion relations traditions developed tools which remain relevant today, they did not achieve the full-scale displacement of perturbative approximation methods that some theorists were hoping for. Causal perturbation theory was thus, in a sense, more successful in articulating a middle ground between Heisenberg and Dyson, if only because its ambitions were more modest.

9 Conclusion

While the drive towards causality-based reformulations of relativistic quantum theory was ultimately not as revolutionary as some might have hoped, it undoubtedly left its mark on the way that QFT is understood today. We close by highlighting two morals for contemporary work on these themes in philosophy of physics.

Firstly, our historical analysis shows that the projects of determining the fundamental theoretical principles underlying an area of physics—a paradigmatically foundational problem—and constructing approximation methods capable of elaborating quantitative results—often maligned as a purely "pragmatic" problem—are deeply intertwined with each other. This is clearly manifested in the story of causal perturbation theory, which essentially became a foundational interrogation of the perturbative approximation scheme itself and is further evidenced by our more impressionistic comments about dispersion relations and the search for non-perturbative approximation schemes in the previous section. Causality conditions did not rise to prominence in high-energy physics through a process of dispassionate conceptual analysis, rather they arose through the struggle to construct calculational methods capable of delivering concrete quantitative results. This

chimes with de Olano et al. (2022)'s claims about the need to integrate engagement with approximation methods into interpretative and foundational debates.

Secondly, the discussion of this paper indicates the need to broaden the range of causality properties in need of philosophical attention. The Bogoliubov causality condition is distinct from the notion of microcausality, since it constrains the behaviour of the theory inside the light cone. The relationship between this S-matrix-based causality condition and the more typical causal properties employed to contemporary axiomatic QFT thus cries out for further philosophical analysis. This is further motivated by contemporary approaches like perturbative algebraic QFT, which bring together elements of the causal perturbation theory and axiomatic QFT traditions—for some discussion of the relationship between the microcausality and Bogoliubov causality within this framework see Fraser and Rejzner (2024). Something similar could also be said about the analyticity conditions appealed to the dispersion relations tradition, which remain poorly understood from both a foundational and historical perspective. According to Hasok Chang, one way that historical research can contribute to our current understanding of science is via the "recovery" of forgotten knowledge (Chang 2017). The results of this paper can be understood as illustrating this idea, though we should perhaps emphasise the recovery of theoretical questions rather than knowledge in this context. We suspect that there are many more important insights waiting to be recovered from a historical investigation of 1950s-1960s relativistic quantum theory.

References

Baacke, J., D. Boyanovsky, and H. de Vega (2001). Initial time singularities in nonequilibrium evolution of condensates and their resolution in the linearized approximation. *Physical Review D 63*(4), 045023.

Barany, M. J., A.-S. Paumier, and J. Lützen (2017). From nancy to copenhagen to the world: The internationalization of laurent schwartz and his theory of distributions. *Historia Mathematica 44*(4), 367–394.

Bjorken, J. D. and S. D. Drell (1964). *Relativistic quantum mechanics*. New York: McGraw-Hill.

Blum, A. S. (2017). The state is not abolished, it withers away: How quantum field theory became a theory of scattering. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 60*, 46–80.

Blum, A. S. (2023). *The Decline and Fall of QED*. Cambridge University Press.

Bogoliubov, N. and O. Parasiuk (1957). Über die Multiplikation der Kausalfunktionen in der Quantentheorie der Felder. *Acta Mathematica 97*(1), 227–266.

Bogoliubov, N. and D. Shirkov (1959). *Introduction to the Theory of Quantized Fields*. New York: Interscience.

Bogoliubov, N. N. (1952a). On the Basic Equations of Quantum Field Theory. *Dokl. Akad. Nauk SSSR 81*, 757–760.

Bogoliubov, N. N. (1952b). On a Class of Basic Equations of Relativistic Quantum Field Theory. *Dokl. Akad. Nauk SSSR 81*, 1015–1018.

Bogoliubov, N. N. (1952c). Variational equations in quantum field theory. *Dokl. Akad. Nauk SSSR 82*, 217–220.

Bogoliubov, N. N. (1955). The causality condition in quantum field theory. *Izv. Akad. Nauk SSSR, Ser. Fiz 19*, 237.

Bogolubov Jr., N. N. (Ed.) (1995). *N.N. Bogolubov: Selected Works*, Volume Part IV: Quantum Field Theory. Gordon and Breach Publishers, New York.

Calderón, F. (2024). The causal axioms of algebraic quantum field theory: A diagnostic. *Studies in the History and Philosophy of Science*.

Carazza, B. and H. Kragh (1995). Heisenberg's lattice world: the 1930 theory sketch. *American Journal of Physics 63*(7), 595–605.

Chang, H. (2017). Who cares about the history of science? *Notes and Records: The Royal Society Journal of the History of Science 71*(1), 91–107.

Cushing, J. T. (1990). *Theory construction and selection in modern physics: The S-matrix*. Cambridge University Press.

de Olano, P. R., J. D. Fraser, R. Gaudenzi, and A. S. Blum (2022). Taking approximations seriously: The cases of the chew and nambu-jona-lasinio models. *Studies in History and Philosophy of Science 93*, 82–95.

Dyson, F. J. (1949a). The radiation theories of Tomonaga, Schwinger, and Feynman. *Physical Review 75*(3), 486.

Dyson, F. J. (1949b). The S-matrix in quantum electrodynamics. *Physical Review 75*(11), 1736.

Dyson, F. J. (1952). Divergence of perturbation theory in quantum electrodynamics. *Physical Review 85*(4), 631.

Earman, J. and G. Valente (2014). Relativistic causality in algebraic quantum field theory. *International Studies in the Philosophy of Science 28*(1), 1–48.

Epstein, H. and V. Glaser (1973). The role of locality in perturbation theory. *AHP 19*(3), 211–295.

Fraser, J. D. (2021). The twin origins of renormalization group concepts. *Studies in History and Philosophy of Science Part A 89*, 114–128.

Fraser, J. D. and K. Rejzner (2024). Perturbative expansions and the foundations of quantum field theory. *Forthcoming in European Physics Journal H*.

Fredenhagen, K. and F. Lindner (2014). Construction of KMS states in perturbative qft and renormalized hamiltonian dynamics. *Communications in Mathematical Physics 332*, 895–932.

Goldberger, M. L. (1955). Use of causality conditions in quantum theory. *Physical Review 97*(2), 508.

Haag, R. (1955). On quantum field theories. *Dan. Mat. Fys. Medd 29*(12), 1–37.

Heisenberg, W. (1943a). Die beobachtbaren grössen in der theorie der elementarteilchen. *Zeitschrift für Physik 120*, 513–538.

Heisenberg, W. (1943b). Die beobachtbaren grössen in der theorie der elementarteilchen II. *Zeitschrift für Physik 120*, 673–702.

Heisenberg, W. (1944). Die beobachtbaren grössen in der theorie der elementarteilchen III. *Zeitschrift für Physik 123*, 93–112.

Heisenberg, W. and W. Pauli (1929). Zur quantendynamik der wellenfelder. *Zeitschrift für Physik 56*(1-2), 1–61.

Helling, R. C. (2012). How I learned to stop worrying and love QFT. *arXiv:1201.2714*.

Kirzhnits, D. A. (1994). V PERVYE POSLEVOENNP1E GODY. In A. N. Sissakjan and D. V. Shirkov (Eds.), *Nikolai Nikolaevich Bogoliubov. Pure Mathematician, applied mathematician, physicist*, pp. 108–111. Dubna: Joint Institute for Nuclear Research.

Klein, L. (1961). *Dispersion relations and the abstract approach to field theory.* Gordon and Breach Publishers, New York.

Landau, L. D., A. Abrikosov, and L. Halatnikov (1956). On the quantum theory of fields. *Il Nuovo Cimento (1955-1965) 3*, 80–104.

Lehmann, H., K. Symanzik, and W. Zimmermann (1955). On the formulation of quantized field theories. *Nuovo Cim 1*(205-225), 80.

Medvedev, B. (1994). N. N. Bogolyubov and the scattering matrix. *Russian Mathematical Surveys 49*(5), 89–108.

Pauli, W. (1940). The connection between spin and statistics. *Physical Review 58*(8), 716.

Rivier, D. and E. Stuecklberg (1948). A convergent expression for the magnetic moment of the neutron. *Physical Review 74*(2), 218–218.

Scharf, G. (2014). *Finite quantum electrodynamics: the causal approach*. Courier Corporation.

Stueckelberg, E. (1944). An unambiguous method of avoiding divergence difficulties in quantum theory. *Nature 153*(3874), 143–144.

Stueckelberg, E. and A. Petermann (1953). La normalisation des constantes dans la théorie des quanta. *Helv. Phys. Acta 26*, 499–520.

Stueckelberg, E. and D. Rivier (1950). A propos des divergences en théorie des champs quantifiés. *Helv. Phys. Acta 23*(Suppl III), 236–239.

Stueckelberg, E. C. and T. Green (1951). Elimination of arbitrary constants in the relativistic theory of quanta. *Helvetica Physica Acta (Switzerland) 24*.

Stueckelberg, E. C. G. (1951). Relativistic quantum theory for finite time intervals. *Phys. Rev. 81*, 130–133.

Sukhanov, A. (1963). The problem of "surface" divergences in the bogolyubov method. *SOVIET PHYSICS JETP 16*(4).

Tomonaga, S.-i. (1946). On a relativistically invariant formulation of the quantum theory of wave fields. *Progress of Theoretical Physics 1*(2), 27–42.

Tomonaga, S.-i. (1966). Development of quantum electrodynamics. *Physics Today 19*(9), 25–32.

Weinberg, S. (1995). *The quantum theory of fields*, Volume 2. Cambridge university press.

Wightman, A. S. (1996). How it was learned that quantized fields are operator-valued distributions. *Fortschritte der Physik/Progress of Physics 44*(2), 143–178.