# 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

Forthcoming in *Synthese*, please cite published version

**Abstract**

This paper examines the development of causal perturbation theory, a reformulation of perturbative quantum 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. Otur reconstruction of this forgotten research program offers new perspectives on contemporary debates about relativistic causality conditions and the problem of ultraviolet divergences.

# Contents

# 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, from axiomatic QFT to the analytic S-matrix. Despite this diversity, one also finds an intriguing consensus among 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.

Notions of relativistic causality remain a central concern in contemporary foundational work on QFT—see Earman and Valente (2014) and Calderón (2024) for recent discussions in the philosophical literature. Much of this work has focused on the various causality axioms adopted in algebraic axiomatizations of QFT leaving aside the question of why alternative definitions of causality became so significant in debates about the formulation of relativistic quantum theory in the first place. This paper unearths a lesser-known strand of 1950s high energy theory, the causal perturbation theory program, which we claim sheds light on this issue. Whereas Freeman Dyson's formulation of perturbative QFT had been based on the integration of the Schwinger-Tomonaga evolution equation, Ernst Stueckelberg, Nikolay Bogoliubov and their collaborators developed an alternative derivation of the perturbation series starting from a causality condition. Causal perturbation theory was thus motivated by a desire to precisely articulate the content of the renormalized perturbative expansion which up to that point had been the main source of empirical predictions in relativistic quantum theory. Comparing the development of causal perturbation theory to other causality-based programs of the period, we argue that causality conditions rose to prominence in this period against a backdrop of a search for new quantitative approximation methods, rather than a philosophical debate about the concept of causality.

In addition to these more general methodological lessons, we extract a number of local interpretative insights from our reconstruction of the causal perturbation theory program. Indeed, this paper can be understood as an exercise in what Chang (2017) calls "recovery"—the retrieval of lost knowledge from neglected episodes in the history of science. Here are three noteworthy instances. Firstly, the "boundary divergences", discussed

in section 3.2, are rarely mentioned in successive literature and raise pressing questions about finite time-evolution in QFT. Secondly, the causality condition which Bogoliubov eventually landed on as the foundation of the causal perturbation theory approach, discussed in section 3.3, is quite different in character from the microcausality condition typically appealed to in axiomatic formulations of QFT, raising questions about the relationships between these causality notions which are yet to be explored in the philosophical literature. Finally, the causal perturbation theory approach formed the basis for a novel treatment of perturbative ultraviolet divergences, discussed in section 3.4, which has the potential to significantly impact interpretive debates surrounding renormalization and the ultraviolet "break down" of QFT. All in all, we hope that our reconstruction of the development of causal perturbation theory will have much to offer less historically minded philosophers of physics.

The paper is structured as follows. Section 2 covers the necessary historical and conceptual background, introducing Heisenberg's S-matrix program and Freeman Dyson's derivation of the perturbative expansion, as well as the criticisms that both of these approaches to relativistic quantum theory faced. Section 3 traces the development of the causal perturbation theory program, with a particular emphasis on the search for an appropriate formulation of the causality condition and the distribution theoretic treatment of the ultraviolet divergences problem. Section 4 concludes with some broader reflections on the different methodological paths being taken by the various causality-based programs of the period and what lessons can be extracted for the philosophical literature on causation in physics.

# 2 Background

## 2.1 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, and 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 became evident that, from second-order, the coefficients of this expansion contained integrals which diverged when the space-time arguments of the summand coincided, or in momentum space where 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. (Oppenheimer

[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 2.2).

1950, p. 276)

The idea here seems to be that 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 theory's 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—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 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 $U_\Lambda$ 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

[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.

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. 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 conditions entered the picture: adding a causality condition to unitarity and Lorentz invariance was a way to add additional dynamical structure without committing oneself to a Hamiltonian field equation.

Why were these new conditions called causality conditions? The conceptual basis of these conditions, and their relationship (if any) to the philosophical literature on causation in science, is a theme we will come back to throughout this paper (and discuss more systematically in section 4). From a historical point of view, however, the use of causality language seems to have originated with Ernst Stueckelberg and the causal perturbation theory program we focus on in this paper. Already in the 1940s, Stueckelberg had suggested that unitarity and Lorentz covariance needed to be supplemented with an additional requirement on the S-matrix which he dubbed a causality condition and interpreted as ruling out retrocausal (i.e. future to past) processes (we analyse Stueckelberg's original causality condition in detail in section 3.1).[3] While these ideas had limited direct influence they appear to have indirectly seeded the broader adoption of causal language in high energy theory.

In particular, the notion of microcausality, which remains the best-known causality condition in relativistic quantum theory today, apparently acquired its name through interactions with Stueckelberg's original formulation of causal perturbation theory. The

[3]The first explicit mention of "the condition that expresses causality" is in (Rivier and Stueckelberg 1948). On earlier uses of "causality" in Stueckelberg's writings, see (Blum 2017, pp. 34–36).

requirement that field operators (and later other classes of operators)[4] associated with space-like separated regions commute, i.e. for a scalar field $\phi(x)$

$$[\phi(x), \phi(y)] = 0, \text{ if } (x-y)^2 < 0, \tag{2}$$

was first introduced by Wolfgang Pauli in his original proof of the spin-statistics theorem. Originally simply writing:

> 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)

Pauli only came to refer to this as a causality condition when comparing it (critically) with Stueckelberg's.[5] From there causality language seems to have become widely adopted. Early axiomatic QFT (Haag 1955; Lehmann et al. 1955) and dispersion theory (Gell-Mann et al. 1954) both added microcausality conditions to Heiseberg's principles of unitarity and Lorentz invariance, by which time the "causality" nomenclature had become standard.[6]

While it is an exaggeration to say that physicists of this period paid no attention to the interpretative status of these conditions, a key claim of this paper is that they were more often prized for what one could do with them rather than their philosophical merits. The real attraction of these conditions was that they represented a way of adding

[4]Later on conditions which require that other sets of operators associated with space-like separated regions, such as the $S(g)$ operators discussed in section 3.3, to commute were also referred to as microcausality conditions. Strictly speaking then, we ought to distinguish different microcausality conditions that range over different sets of local operators.

[5]On 3 August 1948, Pauli wrote, in a letter to Stueckelberg's PhD student Dominique Rivier: "I also don't know what you mean by a 'causality principle.' (It is well known that there are hardly two people who mean the same thing by it – especially not among philosophers.) [...] I formulated the 'causality principle' in quantized field theories as follows: 'all operators that are functions of space-time and can be assigned physically sensible properties on the basis of the current theories (charge densities, electromagnetic field strengths) shall commute for two space-like separated points" (von Meyenn 1993, pp. 552–554).

[6]In Gell-Mann et al. (1954), microcausality is viewed as a relativistic generalization of a condition first used by Ralph Kronig to derive the equations connecting the index of refraction and the absorption coefficient in X-ray dispersion, now known as the Kramers-Kronig relations (Hans Kramers (1928) provided an alternative derivation that did not use a causality-type condition). Kronig (1942) described this condition as "the natural requirement that an electromagnetic field, vanishing [...] for all times $t < 0$ and beginning to act only thereafter, cannot cause the emission of scattered waves before the time $t = 0$." This was first referred to as a "causality condition" by Schützer and Tiomno (1951). So, while this may be an independent origin of the use of the "causality" nomenclature, it arguably postdates Stueckelberg and certainly made the connection to microcausality at a later point. The spread of causal language in this period stands in need of closer investigation.

dynamical structure without adopting a differential evolution equation: thus they offered the prospects of finding a middle ground between Heisenberg's original S-matrix formalism and Dysonian perturbation theory. While Pauli's microcausality condition required adding back the field operators, a departure from Heisenberg's original vision of a pure S-matrix theory, it still allowed dynamical restrictions to be imposed without having to solve, or indeed even formulate field equations. Heisenberg's idea that one ought to start from globally imposed conditions rather than a differential evolution equation in relativistic quantum theory thus lived on in axiomatic QFT and dispersion theory.

Causal perturbation theory also follows this general plan of supplanting a time-evolution equation with a causality condition. There are a number of reasons why it is particularly deserving of the close attention we give it in this paper. As we intimated above, historically it was the first program to introduce a causality condition into relativistic quantum theory, so it remains an obvious starting point for tracing the spread of these ideas. It was also different from axiomatic QFT and dispersion theory in an important respect: whereas these more ambitious programs tried to move away from perturbation theory entirely, causal perturbation theory ultimately developed into an ameliorative reformulation of renormalized perturbation theory. As perturbation theory remains our main source of quantitative information about interacting models, causal perturbation theory arguably retains a direct relevance to contemporary discussions of the foundations of QFT, and since it has been mostly forgotten, studying it offers a number of fresh insights. In particular, the causality condition ultimately formulated by Bogoliubov is not equivalent to the more familiar microcasuality condition and raises novel interpretative questions.

We will return to the relationship between causal perturbation theory, Heisenberg's S-matrix theory and the broader family of causality-based programs it spawned in section 4 after we have investigated the development of causal perturbation theory in detail. To further prepare the way, the following section examines Dyson's derivation of the perturbative expansion; in order to understand the causal perturbation theory approach it will be essential to understand how it differs from this more conventional formulation of renormalized perturbation theory.

## 2.2 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 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 formulation of the perturbative approximation scheme. This section reviews the

derivation of the new renormalized perturbation series found in Dyson (1949a,b). This formulation, which we call Dysonian perturbation theory for short, represents the foil against which causal perturbation theory developed.[7] We pay special attention to how two causality properties enter into Dyson's derivation: firstly, we point out how the microcausality condition introduced in the previous section is incorporated into the Schwinger-Tomonaga equation Dyson starts with; secondly, we highlight how the iterative integration of this equation automatically leads to the time-ordering of operators appearing in the perturbative expansion coefficients, delivering a blueprint for the causality condition adopted in causal perturbation theory.

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

$$id/dt|\psi(t)\rangle = H_I|\psi(t)\rangle, \tag{3}$$

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 (3) still singles out 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:

$$i\frac{\delta\psi(\sigma)}{\delta\sigma(x)} = \mathcal{H}_I(x)\psi(\sigma). \tag{4}$$

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).

[7] We chose the name "Dysonian" rather than something like "conventional" or "orthodox" perturbation theory to distinguish Dyson's derivation from another derivation of the perturbative expansion starting from the path integral expression for the partition function and proceeding via the LSZ reduction formula. The basic form of this derivation can, in fact, be found in the appendix of Lehmann et al. (1955), but it (along with path integral methods more generally) seems to have become widely known much later with the rise of non-abelian gauge theories. The path integral derivation of the perturbative expansion does not seem to have been important for the historical story we tell in section 3 then, though how it relates to the conceptual issues treated in this paper is a question which calls for further investigation.

For all its elegance, however, the Schwinger-Tomonaga equation is somewhat vague. To really make sense of it one needs to specify what sort of variations of the hypersurface $\sigma$ are implied by the variational derivative, and it is in this context that the importance of a microcausality condition becomes evident. Tomonaga (1946) proposed reading (4) 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:

$$[\mathcal{H}_I(x), \mathcal{H}_I(y)] = 0, \text{ if } (x-y)^2 < 0. \tag{5}$$

Tomonaga thus viewed microcausality (applied specifically to the Hamiltonian density operator)[8] to be a necessary condition for the integrability of the Schwinger-Tomonaga equation.

For Dyson, this microcausality condition does not just ensure the integrability of the Schwinger-Tomonaga equation in principle; in a very direct way it enables the iterative integration of the equation, leading to a perturbative expansion for the S-matrix. Explicitly adopting the microcausality of the Hamiltonian density operator as an assumption, Dyson uses it to argue that $\mathcal{H}_I(x)$ depends only on the point $x$ and not on the surface $\sigma$ dyson1949b. In virtue of this locality property, time evolution from an initial to a final surface depends only on an integral over the space-time volume between them. One can thus freely decompose it into a product of time-evolution operators associated with sub-volumes. Introducing a one-parameter family of pre-selected space-like surfaces, Dyson thus proposes writing the interaction picture time-evolution operator as an infinite product of infinitesimal intermediate evolutions:

$$U(t_f, t_i) = \left(1 - i\int_{\tau_1}^{t_f} H_I(\tau)d\tau\right)\left(1 - i\int_{\tau_2}^{\tau_1} H_I(\tau)d\tau\right)\cdots . \tag{6}$$

Here $t_f$ and $t_i$ label the final and initial surfaces respectively, and $\tau_1, \tau_2, ...$ are the labels of the intermediate surfaces. As this equation suggest these labels are in fact all we need

[8] As was noted in footnote 4, one can distinguish different microcausality conditions applied to different sets of operators. Note, however, that if field operators associated with space-like separated points commute then Hamiltonian density operators, which are themselves products of field operators, will also commute at space-like separations.

to keep track of, effectively acting like "times".[9]

It is through trying to evaluate this product that the perturbative expansion and the time-ordering property enter the picture. If the Hamiltonian were time-independent, as in the Schrödinger picture, one could simply take the Hamiltonian $H$ out of the integrals and combine them into a single integral over the entire interval:

$$U(t_f, t_i) = \exp\left\{-iH \int_{t_i}^{t_f} d\tau\right\} = e^{-iH(t_f - t_i)}. \tag{7}$$

When the Hamiltonian is time-dependent, however, 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:

$$U(t_f, t_i) = T_\theta\left(\exp\left\{-i \int_{t_i}^{t_f} H_I(\tau) d\tau\right\}\right) = \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \int_{t_i}^{t_f} dt_1 ... \int_{t_i}^{t_f} dt_n T_\theta[H_I(t_1)...H_I(t_n)]. \tag{8}$$

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

$$T_\theta[H_I(t_1)...H_I(t_n)] = \sum_{p \in P_n} H_I(t_{p_1}) H_I(t_{p_2})...H_I(t_{p_n}) \theta(t_{p_1} - t_{p_2}) \theta(t_{p_2} - t_{p_3})...\theta(t_{p_{n-1}} - t_{p_n}) \tag{9}$$

where the sum is over all permutations $p$ of $n$ variables, and $\theta$ is the Heaviside step function. Roughly speaking, the Heaviside functions 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-ordering operation, and operator products appearing in the Dyson series are referred to as time-ordered products. The reason we have included the (unconventional) $\theta$ subscript is that the causal perturbation theory program will ultimately

[9]Strictly speaking, by introducing a foliation Dyson has already compromised Oppenheimer's demand that covariance be maintained "at all stages" (see the quote in section 2.1). 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 from the start and integrating an interaction picture Schrödinger 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 largely followed their lead on this point. While the improved covariance of Dysonian perturbation theory is often emphasised, more remains to be said about the precise role this played in the breakthroughs of the late 1940s. Maintaining covariance in particular contexts, such as the treatment of longitudinal and transverse components of the electromagnetic field, certainly proved to be crucial for renormalization of perturbative QED, but one might argue that Oppenheimer's demand that covariance be maintained "at all stages" proved to be overblown.

put forward an alternative definition of the time-ordering operation (which we call $T_B[...]$ after Bogoliubov) that does not employ Heaviside functions.

Because time-ordering will be such a crucial concept for our later discussion of causal perturbation theory it is worth immediately making some clarificatory remarks. First of all, it might be worried that since "earlier than" and "later than" are not relativistically invariant locutions time-ordering is a problematically non-relativistic notion. This apparent conflict with covariance can be dissolved however: while some local Hamiltonian density operators will be ordered differently in different foliations by $T_\theta[...]$, such operators will always be space-like separated and thus commute anyway due to microcausality, so their ordering is irrelevant. This is just another way of seeing how microcausality ensures the equivalence of different ways of carving up the Schwinger-Tomonaga evolution using different intermediate surfaces. It is true, however, that (9) is not a covariant expression, raising the question of whether it is possible to express the time-ordering operation in a more relativistic way. This is a more pressing question within the causal perturbation theory program, since the basic idea of this approach is to elevate time-ordering to a fundamental principle; we return to it in section 3.3, where we see that Bogoliubov's time-ordering causality condition can be made fully relativistic.

A second potential concern about the time-ordering operation is that it appears to introduce a distinction between past and future that is not present in the dynamical equations of QFT. Again, though, this is not really the case. If we replace the time ordering in equation 8 with inverse time ordering and switch the initial and final times, we simply get the inverse time evolution operator, which describes evolution backwards in time. In other words, the direction of time ordering in our equations is solely determined by the direction of time evolution we are interested in—in full equivalence with the possibility of taking a differential equation to evolve some initial state both forwards and backwards in time. Furthermore, the appearance of time-ordered products in the perturbation series coefficients is fully consistent with the time-reversal invariance of the S-matrix.

Having said this, time-ordering was connected by the architects of the causal perturbation theory approach to the possibility of reading one's theory as describing time-directed processes. As we shall discuss in the next section, Stueckelberg viewed non-time-ordered products in the series expansion of the S-matrix as allowing for "acausal" processes in which a particle is created at a later time and annihilated at an earlier time. It was through this sort of reasoning that the time-ordering of operators in one's theory came to be understood as a causality property. A final clarifatory point here is that if we are to consider time-ordering as a causality property, it is clearly inequivalent to, and indeed stronger than, microcausality. Microcausality makes a statement about how products of operators associated with space-like points behave—i.e. it tells us about what happens

outside of the light cone. The requirement that operators associated with time-like points form time-ordered products is therefore an additional claim about the structure of the theory inside the light cone. This is reflected in the fact that, whereas Pauli motivated microcausality as a prohibition on superluminal signals, Stueckelberg and Bogoliubov sometimes characterise their causality condition as a prohibition on retrocausal signals.

As already mentioned, the conceptual basis of this causality talk was often quite fuzzy. Rather than fixating on how these conditions relate to causal concepts, we suggest asking the following question: is it possible to identify mathematical conditions which are satisfied in Dysonian perturbation theory and might be usefully abstracted from it? This is the question driving reformulation efforts in the 1950s. As we saw above, a microcausality condition was presupposed by Dyson and was essentially built into the Schwinger-Tomonaga equation: it therefore made sense to ask how much work could be done with microcausality alone. By contrast, the time-ordering of the operator products appearing in the series coefficients was a derived property in Dysonian perturbation theory, following, as we saw from the iterative integration of the Schwinger-Tomonaga equation. Still, it made sense to ask, could we instead treat time-ordering as an axiom? The causal perturbation theory approach can be thought of this way: it promotes time-ordering to a general principle and uses it to derive the form of the series expansion.

Taking the initial and final surfaces to temporal infinity (that is taking the limits $t_f \to \infty$, $t_i \to -\infty$) Dyson obtained a series expansion for Heisenberg's S-matrix, the terms of which can be calculated using Feynman diagrams. Dyson's improved formulation of the perturbative expansion did not automatically solve the ultraviolet divergence problem, however. Starting at second order, the integrals over time-ordered products appearing in the Dyson 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 theory's 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 the first few terms of Dyson's series led to approximations of observables such as the anomalous magnetic moment of the electron which agreed extraordinarily well with experiment.

Dysonian perturbation theory was undoubtedly a major advance over both Heisenberg-Pauli QED and Heisenberg's S-matrix theory from a computational perspective. Despite this success, as we move into through the 1950s the community assessment of Dysonian perturbation theory became increasingly negative. Perhaps the most important reason for this turn of fortune was that the perturbative approximation scheme itself seemed to

many to have reached the limits of its usefulness. The 1950s saw the emergence of new empirical and theoretical problems which a series expansion in the interaction coupling seemed ill-posed to resolve.

On the empirical side, attempts to model nuclear interactions along the lines of perturbative QED floundered, and there seemed to be an in-principle reason for this: the perturbative strategy was based on the assumption that interactions are weak but the strong nuclear interaction is famously strong. It was increasingly felt that new non-perturbative calculational methods would be needed to tame the empirical data coming out of collider experiments. On the theoretical side, the 1950s saw a resurgence of concerns about the inconsistency of QFT.[10] Arguments due to Lev Landau and his collaborators in the Soviet Union suggested that, even if the perturbative ultraviolet divergences could be systematically removed via renormalization, the renormalized charge of the electron still went to infinity at some large finite energy scale—the so-called Landau pole problem. In the early 1950s, Dyson had hoped to prove that renormalized QED perturbation theory converged so that his formalism could be used to demonstrate the existence of solutions of the QED field equations.[11] As it happened, however, he was among the first to argue that, even after renormalization, the perturbation series diverges (Dyson 1952). Again, it seemed that answering the big foundational questions about QFT would mean going beyond the perturbative approximation scheme.

Responding to these fundamental worries about the limits of perturbation theory, programs like axiomatic QFT and dispersion theory hoped to get away from the perturbative approximation entirely. The goal was to find a non-perturbative formulation of relativistic quantum theory which would be capable of meeting these new theoretical and empirical challenges. Causal perturbation theory was, by contrast, somewhat less ambitious. Originating earlier than these more radical reformulation efforts, causal perturbation theory is better thought of as responding to internal problems within Dyson's formulation of the perturbative QFT. Rather than rejecting perturbation theory entirely, it provided a critique, and ultimately a repair, of Dysonian perturbation theory.

One prominent internal issue with Dysonian perturbation theory was the conceptual and mathematical status of the renormalization procedure. Renormalization 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. This led to concerns about the mathematical rigour of the procedure.

[10] See Blum (ming) 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 this controversy which are relevant for the development of causal perturbation theory.

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

The new inconsistency arguments of Landau and others also suggested that renormalization had not really resolved the problems with QFTs high energy behaviour. Renewed worries about the ultraviolet structure of QFT challenged cogency of Dyson's derivation, since positing a differential field 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 Dysonian perturbation theory delivered on this promise since the expansion was divergent. Furthermore, as we shall discuss further in section 3.2, Haag's theorem and the issue of boundary divergences suggested that Dyson's interaction picture evolution operator $U(t_f, t_i)$ does 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 the assumptions about finite time dynamics implicit in his derivation.

Despite its relative obscurity today, we suggest that the causal perturbation theory did make real progress in addressing these internal problems with Dysonian perturbation theory (as we discuss in section 3.4). We are now in position to examine the historical development of this program.

# 3 The Development of Causal Perturbation Theory

## 3.1 Stueckelberg's Causality Condition

To tell the story of causal perturbation theory we need to backtrack a little, as Ernst Stueckelberg's papers which initiated the program actually predate the advent of renormalized QED. While Bogoliubov would eventually repackage his causal derivation of a perturbative expansion for the S-matrix as a more rigorous reconstruction of Dysonian perturbation theory, Stueckelberg conceived it as a novel solution to the problems facing relativistic quantum theory in the early 1940s. Indeed, Stueckelberg presented his ideas as a development of Heisenberg's S-matrix program.[12] He departed from Heisenberg's original visions in two respects: firstly, he was concerned almost exclusively with the construction of a perturbation series for the S-matrix (which Heisenberg's S-matrix papers had tried to avoid), and secondly, he was the first to argue that a causality condition ought to be added to Heisenberg's principles of unitarity and Lorentz invariance. As we will see, the formulation of this causality condition evolved as the causal perturbation theory approach developed, but let us start with Stueckelberg's original intuition.

In Heisenberg's scheme, the unitarity and Lorentz invariance of the S-matrix was ensured

[12] 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).

by writing it as the complex exponential of some hermitian Lorentz scalar $\eta$. An obvious ansatz suggested by Heisenberg is to identify $\eta$ 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 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 "acausal" (Stueckelberg 1944, 144). 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.[13]

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 terms from a series expansion for the S-matrix. 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" (nowadays more commonly known as the "Feynman propagator") which for a scalar field $\phi(x)$ is:

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

[13] 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.

Here $\omega(\mathbf{k})$ is the positive energy $\sqrt{m^2 + \mathbf{k}^2}$ belonging to the momentum vector $\mathbf{k}$, $\theta$ is the Heaviside function, and $x_0$ and $y_0$ are the time components of the four-vectors $x$ and $y$, respectively. $D_c$ 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, representing the propagation of energy from the past to the future. Note that $D_c$ is just a time-ordered product of free field operators at the points $x$ and $y$, so Stueckelberg's condition effectively ensures that the operators appearing in the perturbative coefficients are time-ordered.

Given a first-order term as an input, the requirement that higher-order terms can be decomposed into "causal" propagators, together with unitarity, actually fixes the form of the higher-order terms. It was ultimately argued by Stueckelberg and his collaborators that what one gets by following this recipe agrees with the 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 could have provided an alternative basis for the development of renormalized QED in the post-war period. As it happened, however, it was largely ignored. The most important reason for this was likely bad timing; by the time Stueckelberg was capable of delivering any concrete results, the famous second-order QED calculations of Schwinger and Feynman were widely known, and Dysonian perturbation theory had stolen his thunder.

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 popularised by Dyson. 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 3.4, 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 phase, 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 Bogoliubov as a foundationally motivated reconstruction project rather than a novel theory. Bogoliubov introduced a clearer and better-motivated causality condition and worked out its consequences more systematically. Before we get to Bogoliubov's, however, we will first recount

another Stueckelbergian idea: the appearance of a novel class of "boundary" divergences in the perturbative expansion for finite (i.e. non-asymptotic) times. In addition to their importance for the development of causal perturbation theory, these divergences present a largely forgotten foundational problem with Dysonian perturbation theory which it seems to us is worthy of "recovery".

## 3.2 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);[14] 2) the more serious *boundary divergencies* (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 3.4. The second type of divergence, however, is likely unfamiliar even to QFT aficionados. As it happened, the analysis of these "boundary divergences" led (in a rather convoluted way) to Bogoliubov'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).[15] 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

[14] On the connection between the work of Stueckelberg's student Dominique Rivier and the more well-known Pauli-Villars regularization method, cf. Schweber, QED.

[15] 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.

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 an expression for the asymptotic S-Matrix (whether derived from a time-evolution equation or a causality condition) 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 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:

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

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:

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

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 $\omega$ leading to a divergent expression for the matrix element.[16]

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.[17] 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;[18] 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

[16] If one takes the renormalized expression for $\Sigma$ at second order in perturbation theory from, e.g., Weinberg (2002) Eq. 11.4.14, and takes the limit of large $\omega$ (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.

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

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

renormalization procedure which cures the usual ultraviolet divergences does nothing to alleviate these new infinities. Bogoliubov 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).[19] 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 Schwinger-Tomonaga 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 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, triggering 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. From the late 1940s, Nikolay Bogoliubov was reading the latest Physical Review papers on renormalized QED with his group at the

[19]There is a puzzling discrepancy between Bogoliubov's and Stueckelberg's discussions of boundary divergences. The Bogoliubov 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.

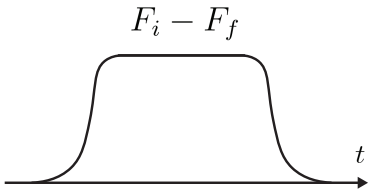


Fig 1

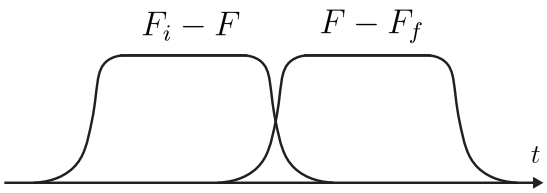


Fig 2

Steklov Institute of Mathematics in Moscow (Medvedev 1994).[20] 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, rebranding it as a mathematical reconstruction project and putting forward an improved formulation of the causality condition.

## 3.3 Bogoliubov's Causality Condition

Arriving on the scene after the empirical successes of Dysonian perturbation theory, and with a background in mathematics, Bogoliubov approached QFT from a different perspective than Stueckelberg. Now the dust had settled somewhat, the question for theorists of his orientation was whether it was possible to make good mathematical sense out of the advances of the late 1940s.[21] What initially captured Bogoliubov's attention, it seems, was the fundamental issue of how the dynamics of a relativistic quantum theory ought to be represented. 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 this question.[22] He agreed with Stueckelberg (1951) that the boundary divergences necessitated building a smooth switching function into one's 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

[20]Indeed, Bogoliubov's group appears to have been the first to broach renormalized QFT in the Soviet Union, predating the Landau group's engagement with the subject (Kirzhnits 1994).

[21]We can see Bogoliubov as part of a larger move towards more mathematically rigorous work on QFT in this period, alongside figures like Rudolf Haag and Arthur Wightman. As we discuss in section 4 what distinguished Bogoliubov from the axiomatic field theorists was his attitude towards the perturbative expansion.

[22]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.

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 the notion of time-ordering causality that Stueckelberg had been gesturing at, 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 a generalized $S(g)$ evolution operator. Now, as we saw in section 2.2, 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 expansion.[23]

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 $g(x)$:

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

[23]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.

Taking $g(x)$ to be Heaviside functions centred on space-like hypersurfaces would take 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.[24] 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.[25] 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 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 a causality condition.

While Bogoliubov now embraced the primacy of the causality condition he found Stueckelberg's attempts to formulate such a condition to not be "sufficiently clear and general", explaining why his "ideas did not receive wide recognition" (Bogoliubov 1955, 237). One

[24] 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.

[25] 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.

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 ultimately put forward two versions of his new causality condition using the switching function: 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. Bogoliubov reasoned that the infinitesimal change in the wave function associated with an infinitesimal change in the switching function at a point $x$ should only depend on the shape of the switching function in the future lightcone of $x$. 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:[26]

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

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 formulation of his causality condition (Bogoliubov and Shirkov 1959). We focus on this version here for presentational purposes as we find the derivation of the perturbative

[26]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.

expansion from it easier to grasp.[27] To motivate the integral version of Bogoliubov's 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, Bogoliubov'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:

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

Intuitively, this statement says that, if the S-matrix is factorized into a product of operators associated with sub-sections of the evolution it forms a time-ordered product, with the operator associated with the earlier space-time region acting on the wave function first. We shall return to the interpretation of this condition shortly, but it will be useful to start by sketching how it can be used to derive the form of the perturbative expansion.[28]

With the introduction of the switching function, a general series expansion for the S-matrix takes the form:

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

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 Bogoliubov's integral causality condition and rearrange the resulting terms one can obtain two equivalent series:

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

[27] Furthermore, it is the integral version of Bogoliubov's causality condition that one finds in later mathematical physics work on the causal perturbation theory approach. See Epstein and Glaser (1973).

[28] The derivation of the perturbative expansion given here follows the presentation of Scharf (2014) and is thus anachronistic but has been preferred here for reasons for compactness.

Where none of the the time components of $\{x_{m+1}, ..., x_n\}$ are in the future lightcone of any of the points $\{x_1, ..., x_m\}$. 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 $S_1(x) = iH_I(x)$, which amounts to a choice of interaction term for one's perturbative model, one obtains higher-order terms of the form:

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

where $T_B[...]$ is the time-ordering operation (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 density. If one takes the limit $g(x) \to 1$, returning to the usual asymptotic S-matrix operator, one indeed obtains the form of the Dyson series.

Whereas the time-ordering of the series coefficients in Dyson's derivation followed from the integration of the Schwinger-Tomonaga equation, what Bogoliubov had essentially demonstrated is that it is possible to instead treat time-ordering causality as a primitive posit and derive the expansion. One immediate advantage of this alternative derivation is that one could sidestep the problems with the Schwinger-Tomonaga equation and Dyson's interaction picture evolution operator by simply not employing them. Another advantage of this approach was that it allowed for a more careful treatment of the time-ordering property. The implementation of the time-ordering operation which features in Bogoliubov's derivation ($T_B[...]$) is not in fact identical to the Heaviside time-ordering operation which appears in Dyson's derivation ($T_\theta[...]$). For one thing, Bogoliubov had succeeded in expressing time-ordering in a fully relativistic fashion (answering a question raised in section 2.2). More importantly, however, because Bogoliubov's causality condition left the behaviour of the time-ordered products at coincident space-time points unspecified it allowed one to avoid writing down divergent integrals. As we shall see in the next section, this led to a new analysis of the ultraviolet divergences problem within the causal perturbation theory approach.

Before we get to that though we close this section with some preliminary remarks about the physical interpretation of Bogoliubov's causality condition. One issue we identified with Stueckelberg's justification for his condition was his reliance on a controversial physical reading of virtual processes associated with individual terms in the series expansion. Bogoliubov's causality condition by contrast, is initially formulated in non-perturbative terms, holding out the prospects of a more robust physical interpretation. Challenges still loom here, however. Bogoliubov took his condition to express the requirement that "any

event occurring in the system may exert an influence on the evolution of the system only in the future and cannot exert any influence on the behavior of the system in the past at times preceding the given event" (Bogoliubov and Shirkov 1959, 200-201). However, reading his condition this way seems to require attaching a questionable physical significance to the $S(g)$ operators and the switching function itself. Since one typically takes $g \to 1$ at the end of calculations, with $S(g)$ acting as a mere calculational intermediaries, their meaning even at an operational level is far from clear. As a result, later mathematical physics work on perturbative QFT would often simply justify Bogoliubov's causality condition by its implications—Scharf writes "the correctness of the [causality condition] can only be shown by working out its consequences"(Scharf 2014, 163), for instance.

At the very least then, it is clear that Bogoliubov's causality condition raises novel interpretative questions that stand in need of a more careful philosophical analysis.[29] One obvious remaining question, for instance, is whether Bogoliubov's causality condition conflicts with retrocausal interpretations of quantum theory.[30] While this topic requires further investigation, we suspect that it does not. If Bogoliubov causality says anything about retrocausal influence it is something like the following: operations at some point $x$ do not affect the statistics of measurements in the past lightcone of $x$. Retrocausal solutions of the measurement problem do not require this, however. This is analogous to the fact that microcausality does not in fact rule out the possibility of non-local hidden variable theories. A final point we would like to make before we move on is that Bogoliubov's condition helps to clarify the relationship between time-ordering causality and microcausality. Returning to the integral causality condition, if we also have that none of the points in $G_2$ are in the future light cone of any of the points in $G_1$, i.e., if $G_1$ and $G_2$ are space-like separated, we get the additional condition $S(g_1 + g_2) = S(g_1)S(g_2)$, so $S(g_1)$ and $S(g_2)$ commute. Thus, (23) implies a version of the microcausality property (applied to the $S(g)$ operators).[31] This precisely expresses the intuitive distinction between microcausality and time-ordering made in section 2.2: Bogoliubov causality strengthens microcausality by making a claim about how products of local operators behave inside the lightcone as well as outside of it.

## 3.4 Rethinking the Ultraviolet Divergence Problem

Arguably, the most important achievement of the causal perturbation theory programs was that it produced to a novel formalization of the ultraviolet divergences problem and the perturbative renormalization procedure. Recall that in Dysonian perturbation theory

[29]Further impetus to analyse Bogoliubov's condition comes from its role in influential contemporary programs like perturbative algebraic QFT. See Fraser and Rejzner (2024) for a recent discussion.

[30]See Friederich and Evans (2023) for a review of retrocausal approaches in quantum foundations.

[31]We would like to thank Kasia Rejzner for pointing this out to us.

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 relationships between distinct 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. Bogoliubov 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 Bogoliubov called 'quasi-local operators', $\Lambda_n$, 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.[32] Distributions are objects which generalize the standard notion of a function. It is possible to uniquely associate a

[32]Schwartz's work on distributions in fact ran 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 33.

locally integrable function, $f(x)$, with a functional that takes test functions, $g(x)$, to the numbers:

$$\mathcal{T}_f : g(x) \to \int_{-\infty}^{\infty} f(x)g(x)dx; \tag{27}$$

The basic idea of distribution theory is to consider a larger class of functionals that includes more singular objects, such as the Dirac delta "function",

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

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.[33] 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)

[33]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).

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.

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.[34] 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 $x - y = 0$. 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 2.2, the Heaviside functions in the conventional time-ordered product, $T_\theta[...]$, arise automatically from the integration of the Schwinger-Tomonaga equation performed by Dyson.[35] 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)

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

[35] More precisely, they arise from the limits of the integrals in equation (6), and thus from the space-like hypersurfaces employed by the Schwinger-Tomonaga equation.

The causality condition fixes the behaviour of $S_n$ everywhere except at coincident space-time points; it thus allows us to construct a time-ordered product on a space of test 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.[36] 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 rig-

[36]One can also construct perturbation series in quantum mechanics using a version of Bogoliubov's causality condition rather than the Schrödinger 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.

orous resolution of the ultraviolet divergences problem. From the perspective of this later mathematically mature articulation of causal perturbation theory, ultraviolet divergences 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 (and will be examined more carefully in forthcoming work (Fraser and Miller 2025)), 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.

# 4 Reflections on Causality Conditions in Physics

Having now explored the development of the causal perturbation theory program in some detail, we conclude this paper with some broader reflections on the rise of causality conditions in high energy theory.

Let us start by addressing a question which may have been on less historically minded readers' minds for some time: what does all this have to do with causation? The use of the term "causality" in relativistic quantum theory contradicts Bertrand Russell's notorious claim that causal language is absent from physics (Russell 1912), leading to some controversy in the recent literature on causation in physics. Some philosophers, such as Norton (2003), take a deflationary view of causality conditions in physics, arguing that "causality" acts as an empty honorific, with little connection to substantive notions of causation. On the other hand, if we adopt Ben-Menahem's characterisation of a causal principle as "any constraint that delimits change" (Ben-Menahem 2018, 15), then the causality conditions discussed in this paper will clearly count as genuinely causal, though perhaps in

a fairly trivial way since unitarity and Lorentz invariance will presumably also be classed as causal principles on this view. An approach to advancing a more selective connection between the interventionist approach to causation and causality conditions in physics has recently been advanced by Weinberger et al. (2023). They suggest that some of these conditions might be viewed as part of the "worldly infrastructure" of causation—roughly speaking, constraints which must be satisfied for interventionist style causal modeling to be possible.

We do not think that the historical data examined in this paper offers decisive support to any of these philosophical views of causality conditions in physics (whether Bogoliubov's causality condition can be understood as encapsulating "worldly infrastructure" of causation, for instance, seems to us to be an open question). However, it does offer an important correction to the way this debate has been framed. A key moral which emerges from our discussion is that physicists of the 1950s were not primarily concerned about the philosophical status of the conditions they were writing down; rather, they evaluated these conditions according to their perceived utility in achieving concrete theoretical goals.

As we suggested in section 2, the overarching goal of many programs of 1950s high energy theory was to articulate a middle ground between Heisenberg's S-matrix theory and Dysonian perturbation theory. It was clear that Heisenberg's principles of unitarity and Lorentz invariance were too minimal a starting point for relativistic quantum theory, but the richer dynamical structure posited by Dyson appeared to have added too much. Doubts about the existence of solutions of the Schwinger-Tomonaga equation intensified in the 1950s, and 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 2.2, the Schwinger-Tomonaga framework incorporated properties such as microcausality and perturbative time-ordering that did not follow from Heisenberg's principles. Functionally, at least, "causality" referred to any dynamical principle which was stronger than unitarity and Lorentz invariance but weaker than a Hamiltonian evolution equation. Abstracting these properties from the Dysonian framework and working with them directly offered an enticing way forwards for high-energy theorists in this period.

Causal perturbation theory was an especially clear instantiation of this basic strategy. We would like to highlight some major differences between causal perturbation theory and other causality-based programs, however, which sheds further light on the broader adoption of these conditions (these remarks will necessarily be somewhat speculative as the historical development of these programs is yet to be carefully studied). The methodological path taken by causal perturbation theory was a comparatively conservative one. What Bogoliubov essentially did was start with the conventional perturbative approxi-

mation scheme and ask whether the same results could be derived from a weaker set of assumptions. This meant that, while he was arguably able to improve the internal coherence of perturbative QFT, he was not able to overcome any of its intrinsic limitations: Causal perturbation theory did nothing to address the divergence of the expansion since it reproduces the large-order behaviour of the Dyson series; and while it offered an ameliorative treatment of the perturbative ultraviolet divergences, it left the Landau pole problem untouched, suggesting that QED breaks down anyway for different (non-perturbative) reasons. 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 hope that finding a stable middle ground between Heisenberg and Dyson would also lead to new calculational resources and an escape from the perturbative approximation scheme which seemed to inevitably fail in the context of strong nuclear interactions.

The need to go beyond perturbation theory in order to finally slay QFT's foundational 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 an evolution equation. In fact, the main role played by the field operators in the LSZ formalism is to implement a principle of microcausality 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 procedures which Dysonian perturbation theory remained tied to.

This optimism about the possibility of extracting non-perturbative information from causal-

ity 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 analyticity 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 arguably more successful in articulating a middle ground between Heisenberg and Dyson, if only because its ambitions were more modest. While this look at the broader landscape of causality-based reformulations reveals a healthy amount of methodological diversity it also hammers home a more general moral: 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. It is further evidenced by our more impressionistic comments about dispersion relations and the search for non-perturbative approximation schemes above. Causality conditions emerged out of a struggle to construct calculational methods capable of delivering concrete quantitative results rather than a dispassionate debate about the semantics of causal concepts, and this

ought to be taken into account in their philosophical reception. This final point chimes with Ruiz de Olano et al. (2022)'s claims about the need to integrate engagement with approximation methods into interpretative and foundational debates.

# Acknowledgements

Conversations with Michael Miller played a crucial role in the early stages of the long research project which eventually produced this paper. Kasia Rejzner, Michael Dütsch, and Günter Scharf very kindly corresponded with us on the later mathematical physics implementations of the causal perturbation theory approach. Porter Williams, Francisco Calderón and two anonymous referees gave comments on drafts of the manuscript, leading to many improvements. James Fraser would like to acknowledge funding from the ASYMPTOPHYS project (ANR-22-CE540002, hosted by IHPST, CNRS, UMR8590, France).

# 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.

Ben-Menahem, Y. (2018). *Causation in science*. Princeton University Press.

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. (Forthcoming). *Probing the Consistency of QFT I: From Non-Convergence to Haag's Theorem*. 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 104*, 98–108.

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.

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 M. E. Miller (2025). Why are there ultraviolet divergences at all? *unpublished manuscript*.

Fraser, J. D. and K. Rejzner (2024). Perturbative expansions and the foundations of quantum field theory. *The European Physical Journal H 49*(1), 10.

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

Friederich, S. and P. W. Evans (2023). Retrocausality in Quantum Mechanics. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023 ed.). Metaphysics Research Lab, Stanford University.

Gell-Mann, M., M. Goldberger, and W. Thirring (1954). Use of causality conditions in quantum theory. *Physical Review 95*, 1612–1627.

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). В ПЕРВЫЕ ПОСЛЕВОЕННЫЕ ГОДЫ. 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.

Kramers, H. A. (1928). La diffusion de la lumière par les atomes. In *Atti del Congresso Internazonale dei Fisici, 11–20 Settembre 1927*, Volume 2, Bologna, pp. 545–557. Nicola Zanichelli.

Kronig, R. (1942). Algemeene theorie der diëlectrische en magnetische verliezen. *Nederlands Tijdschrift voor Natuurkunde 9*, 402–409.

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.

Norton, J. D. (2003). Causation as folk science. *Philosopher's Imprint 3*(4), 1–22.

Oppenheimer, J. R. (1950). Electron theory. In *Les Particules Élémentaires*, Brussels, pp. 269–286. Institut International de Chimie Solvay: R. Stoops.

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

Rivier, D. and E. C. G. Stueckelberg (1948, July). A convergent expression for the magnetic moment of the neutron. *Physical Review 74*(2), 218 (Erratum p. 986).

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

Ruiz de Olano, P., 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.

Russell, B. (1912). On the notion of cause. In *Proceedings of the Aristotelian society*, Volume 13, pp. 1–26. JSTOR.

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

Schützer, W. and J. Tiomno (1951). On the connection of the scattering and derivative matrices with causality. *Physical Review 83*(2), 249–251.

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.

von Meyenn, K. (Ed.) (1993). *Wolfgang Pauli: Wissenschaftlicher Briefwechsel mit Bohr, Einstein, Heisenberg u.a.*, Volume III: 1940-1949. Berlin: Springer.

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

Weinberger, N., P. Williams, and J. Woodward (2023). The worldly infrastructure of causation. *BJPS, https://doi.org/10.1086/730698*.

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.